The Dissertation Committee for Anush Krishna Moorthy
certifies that this is the approved version of the following dissertation:

Natural Scene Statistics Based Blind Image Quality Assessment and Repair

Committee:

Alan C. Bovik, Supervisor

Wilson S. Geisler

Joydeep Ghosh

Kristen Grauman

Haris Vikalo

Natural Scene Statistics Based Blind Image Quality Assessment and Repair

by

Anush Krishna Moorthy, B.E., M.S.E.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2012

For Amma, Appa and Silky.

Acknowledgments

I write this section barely three days before my defense, for in the lull before the storm I was reminded of the multitude of people who have helped and guided me through these 4-odd years towards my goal.

First in this pantheon (yes, the word is appropriate in this context) is *Amma*, my mother. If there ever was one person that I'd refer to as my rock, it would be her. While we have never found common ground on any of our philosophical ideas, I believe that it is only because of her that I am capable of any coherent thought, for she has not only given me life but also instilled in me the underlying tenaciousness that is the hallmark of any man. While I can ramble on about the contributions of my mother towards every endeavor of my life, coloring it with florid epithets, I shall quote the cliche – her actions are beyond description even by competent wielders of the language, let alone by an amateur fledgling such as myself.

Next up is *Appa*, my father. He is the quiet *yin* to my mother's garrulous *yang*, the *purusha* to her *prakriti*. Any semblance of moral rectitude that I may have is solely due to my father, who is the most upright individual I have met. While this may seem like a handicap in this *kaliyuga*, it is my firm belief that his innocence and taintless character function as a beacon in a world surrounded by darkness – a beacon that I endeavor everyday to match.

My regard for my mother stems from him as well, for I have not seen a more dedicated son. An ideal father, *Appa* has been the intellectual bedrock from which I have sprung forth.

Now onto the Guru. Dr. Bovik, as all of his students would attest, is the world's best advisor; one that graduate students dream of as they are whipped into action by the harsh overlords of the academic world elsewhere. At LIVE, Dr. Bovik wields the baton with a finesse of a concert conductor; never harsh, quick to compliment and slow to criticize, he epitomizes the *Bheeshma* epithet like no other. Allowing us sheep the freedom to explore, while constantly stirring the creative juices without actually leading us to the solution, Dr. Bovik has honed his skills as an advisor to perfection. I have learnt that discipline and passion will take a person a long way in life, and maintaining a calm visage while paddling like a duck beneath the surface is the key to success. As they say, still waters run deep.

The final major acknowledgement belongs to God. While I question his existence and disrespectfully berate the orthodoxy of religious beliefs on a daily basis, there is still a part of me that wonders at the possibility of His existence. If not in the traditional form as an intelligent designer, I am sure He exists in the thread that all creatures dead or alive share. In the form of an all knowing God, or in the form of the *purusha-prakriti* of the Vedas, or in the form of fate or luck or destiny, I believe that He has guided me to this juncture, as He shall in the future. If there is an all knowing God, then I submit my humble apologies at my blasphemies, and if there isn't then I shall still consider

that *thread* as the guiding beam of light that has laid the path upon which I find no thorns. In either case, He requires a big acknowledgement.

Having waxed eloquently on the *mata-pita-guru-deivam* quadrumvirate, I shall not bore the reader with mundane details of the remaining acknowledgees, but instead, shall simply list names. If you know me and do not find mention here, I seek apology and blame my mental ineptitude, for man is inspired by every single person he meets, and to all of the people I know, I extend my sincerest thanks.

LIVE members, past and present: Yang, Sina, Kalpana, Joonsoo, Rajiv, Gautam, Ajay, Anish, Ming, Michele, Che-Chun, Dinesh, Lark.

Friends of yore: Ronak, Sid, Shirish, Shiv, Harshit, Vami, Nisha and their better (bitter) halves.

Roomies, past and present: Pranav, Keith, Surbhi, Manohar, Anish.

Friends, newer and nascent: Abhik, Aditya, Akshay, Aneesh, Anish Anuj, Arundhati, Bade Bhaiya, DK, Gaur, Gill, Guneet, Harpreet, Harsh, Kiran, Kriti, Neha, Nuke, PK, Poolkeshi, Pranav, Praneeth, Preeti, Raghav, Ravindara, Sarabjot, Sharayu, Shatam, Shruti, Siddhartha, Sindu, Tanvi, VDC, Vikram, Vimal, Vishal.

Finally, to my committee and Dr. de Veciana, for their time and in many cases, for their advice and push in the right direction.

Natural Scene Statistics Based Blind Image Quality Assessment and Repair

Publication No. _____

Anush Krishna Moorthy, Ph.D.
The University of Texas at Austin, 2012

Supervisor: Alan C. Bovik

Progress in multimedia technologies has resulted in a plethora of services and devices that capture, compress, transmit and display audiovisual stimuli. Humans – the ultimate receivers of such stimuli – now have access to visual entertainment at their homes, their workplaces as well as on mobile devices. With increasing visual signals being received by human observers, in the face of degradations that occur to due the capture, compression and transmission processes, an important aspect of the quality of experience of such stimuli is the *perceived visual quality*. This dissertation focuses on algorithm development for assessing such visual quality of natural images, without need for the 'pristine' reference image, i.e., we develop computational models for no-reference image quality assessment (NR IQA).

Our NR IQA model stems from the theory that natural images have certain statistical properties that are violated in the presence of degradations,

and quantifying such deviations from *naturalness* leads to a blind estimate of quality. The proposed modular and easily extensible framework is distortion-agnostic, in that it does not need to have knowledge of the distortion afflicting the image (contrary to most present-day NR IQA algorithms) and is not only capable of quality assessment with high correlation with human perception, but also is capable of identifying the distortion afflicting the image. This additional distortion-identification, coupled with blind quality assessment leads to a framework that allows for blind general-purpose image repair, which is the second major contribution of this dissertation. The blind general-purpose image repair framework, and its exemplar algorithm described here stem from a revolutionary perspective on image repair, where the framework does not simply attempt to ameliorate the distortion in the image, but to ameliorate the distortion, so that visual quality at the output is maximized.

Lastly, this dissertation describes a large-scale human subjective study that was conducted at UT to assess human behavior and opinion on visual quality of videos when viewed on mobile devices. The study lead to a database of 200 distorted videos, which incorporates previously studied distortions such as compression and wireless packet-loss, and also dynamically varying distortions that change as a function of time, such as frame-freezes and temporally varying compression rates. This study – the first of its kind – involved over 50 human subjects and resulted in 5,300 summary subjective scores and time-sampled subjective traces of quality for multiple displays. The last part of this dissertation analyzes human behavior and opinion on time-varying video qual-

ity, opening up an extremely interesting and relevant field for future research in the area of quality assessment and human behavior.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Man is a visual animal. Through evolution, man has always been fascinated by what he can see and imagine and has endeavored to re-create this world on canvas - from pre- historic cave paintings to modern-day films accompanied by sounds and visual effects. Although arguments on the reason for cave paintings persist, it is clear that at least in todays world, much of the created visual stimuli are for entertainment or informational purposes.

The urge to capture the world around us has lead to the creation of devices which are increasingly capable of doing so, and burgeoning demand has lead to increasing availability at ever-reducing costs. Transmission, storage and display of visual stimuli have also escalated. At the end of this chain is the receiver - the human observer. It should be obvious that with so many different devices capturing, storing, compressing, transmitting and displaying visual stimuli, the receiver is bound to receive stimuli of varying levels of palatability.

This dissertation deals with visual quality assessment, and aims to understand human opinion on visual palatability/quality and algorithmically capture this palatability. Let us now hold the reader's hand and introduce

the concepts relevant to visual quality assessment.

1.1 Concepts in Quality Assessment

Imagine this situation - you are given two images/videos, both having the same content but one of the images/videos is a 'low quality' (distorted) version of the other and you are asked to rate the low quality version vis-a-vis the original (reference) image/video on a scale of (say) 1-5 (where 1 is bad and 5 is excellent). Let us further assume that we collect a representative subset of the human populace and ask them the same question, and instead of just asking them to rate one pair of images/videos, we ask them to rate a whole set of such pairs. At the end of the day we now have a set of ratings for each of the distorted images/videos, which when averaged across users gives us a number between 1-5. This number represents the mean opinion score (MOS) of that image/video and is a measure of the perceptual quality of the image/video. The setting just described is called subjective evaluation of video quality and the case in which the subject is shown both the reference and the distorted image/video is referred to as a double stimulus study. One could imagine many possible variations to this technique. For example, instead of showing each image/video once, let us show each image/video twice so that in the first pass the human 'decides' and in the second pass the human 'rates'. This is a perfectly valid method of collecting subjective scores and along with a plethora of other techniques forms one of the possible methods for subjective evaluation of image/video quality. Each of these methods is described in a

document from the International Telecommunications Union (ITU) [288] . If only we always had the time to collect a subset of the human populace and rate each image/video that we wish to evaluate quality of, there would have been no necessity for this dissertation or the decades of research that has gone into creating algorithms for this very purpose.

Algorithmic prediction of image/video quality is referred to as objective quality assessment, and as one can imagine it is far more practical than a subjective study. Algorithmic image/video quality assessment (IQA/VQA) is the main focus of this dissertation, although we shall study human opinion on visual quality as well. Before we delve directly into the subject matter, let us explore objective assessment just as we did with the subjective case. Imagine you have an algorithm to predict quality of an image/video. At this point it is simply a 'black-box' that outputs a number between (say) 1-5 - which in a majority of cases correlates with what a human would say. What would you imagine the inputs to this system are? Analogous to the double stimulus setup we described before, one could say that both the reference and distorted images/videos are fed as inputs to the system - this is full reference (FR) quality assessment. If one were to imagine practical applications of FR IQA/VQA, one would soon realize that having a reference video is infeasible in many situations. The next logical step is then truncating the number of inputs to our algorithm and feeding in only the distorted image/video - this is no reference (NR) IQA/VQA. Does this mean that FR IQA/VQA is not an interesting area for research? Surprisingly enough, the answer to this question

is NO! There are many reasons for this, and one of the primary ones is that FR IQA/VQA is an extremely difficult problem to solve. This is majorly because our understanding of perceptual mechanisms that form an integral part of the human visual system (HVS) is still at a nascent stage [245, 303]. FR IQA/VQA is also interesting for another reason - it gives us techniques and tools that may be extended to NR IQA/VQA.

Thinking solely from an engineering perspective one would realize that there exists another modality for IQA/VQA. Instead of feeding the algorithm with the reference and distorted images/videos, what if we fed it the distorted image/video and *some features* from the reference image/video? Can we extract features from the reference video and embed them into the video that we are (say) transmitting? If so, at the receiver end we can extract these reference features and use them for image/VQA. Such assessment of quality is referred to as reduced-reference (RR) image/VQA. This dissertation is mainly concerned with algorithmic approaches for NR IQA and its application.

In describing the RR technique, we have inadvertently stumbled upon the general system description for which algorithms described in this dissertation are designed. There exists a pristine reference image/video which is transmitted through a system from the source. At the receiver, a distorted version of this image/video is received whose quality is to be assessed. Now, the system through which the image/video passes could be a compression algorithm. In this case, as we shall see, measures of blockiness and bluriness are used for NR IQA. In case the system is a channel that drops packets, the effect

of packet loss on quality may be evaluated. We now briefly describe how the performance of an algorithm is evaluated.

1.1.1 Performance Evaluation & Databases

We know that the aim of IQA/VQA is to create algorithms that predict the quality of an image/video such that the algorithmic prediction matches that of a human observer. For this section let us assume that we have an algorithm which takes as input a distorted image/video (and some reference features) and gives us as output a number. The range of the output could be anything, but for this discussion, let us assume that this range is 0-1, where a value of 0 indicates that the signal is extremely bad and a value of 1 indicates that the signal is extremely good. We also assume that the scale is continuous, i.e., all possible real-numbers between 0 and 1 are valid algorithmic scores. With this setup, the next question one should ask is, 'How do we know if these numbers generated are any good?'. Essentially, what is the guarantee that the algorithm is not spewing out random numbers between 0 and 1 with no regard to the intended viewer?

The ultimate observer of a visual signal is a human and hence his perception of quality is of utmost importance. Hence, a set of images/videos are utilized for a subjective study and the perceptual quality of the image/video is captured in the MOS. However, picking (say) 10 images/videos and demonstrating that the algorithmic scores correlate with human subjective perception is no good. We require that the algorithm perform well over a wide variety of

cases, and hence the database on which the algorithm is tested must contain a broad range of distortions and a variety of content, so that the stability of its performance may be assessed. In order to allow for a fair comparison of algorithms that are developed by different people, it is imperative that the IQA/VQA database, along with the subjective MOS be made publicly available. For IQA, currently, the LIVE image quality assessment database is the de facto standard and incorporates a wide range of distortion types and degradation levels [264]. In this proposal, we will not be concerned with IQA database creation and hence we shall cease talking about it here and focus on VQA.

One publicly available dataset for VQA is the popular Video Quality Experts Group (VQEG) FRTV Phase-I dataset [297]. The VQEG dataset consists of 20 reference videos, each subjected to 16 different distortions to form a total of 320 distorted videos. In [297], a study of various algorithms was conducted on this dataset and it was shown that none of the assessed algorithms were statistically better than peak signal-to-noise ratio (PSNR)[1]! Over the years, many new FR VQA algorithms which perform well on this dataset have been proposed [247, 313]. However, the VQEG dataset is not without its drawbacks [255, ?], and hence we (and other researchers) have proposed attractive alternatives to the VQEG database [182, ?].

Now that we have a dataset with subjective MOS and scores from an

[1]Why PSNR is a poor measure of visual quality is described in [94] and [310].

algorithm, our goal is to study the correlation between them. In order to do so, Spearman's Rank Ordered Correlation Coefficient (SROCC) [266] is generally used [297]. SROCC of 1 indicates that the two sets of data under study are perfectly correlated. Other measures of correlation include the Linear (Pearson's) correlation coefficient (LCC) and the root-mean-square error (RMSE) between the objective and subjective scores. LCC and RMSE are generally evaluated after subjecting the algorithms to a logistic function. This is to allow for the objective and subjective scores to be non-linearly related. For eg., figure 1.1 shows a scatter plot between MOS scores from the VQEG dataset and an FR VQA algorithm [173]. As one can see, the two are definitely correlated, only that the correlation is non-linear. Transformation of the scores using the logistic accounts for this non-linearity and hence application of LCC and RMSE make sense. It is essential to point out that application of the logistic in no way constitutes 'training' an algorithm on the dataset (as some authors claim). It is simply a technique that allows for application of the LCC and RMSE as statistical measures of performance. A high value (close to 1) for LCC and a low value (close to 0) for RMSE indicate that the algorithm performs well.

Having summarized how one would analyze a VQA algorithm, let us move on to the human visual system whose properties are of tremendous importance for developing VQA algorithms.

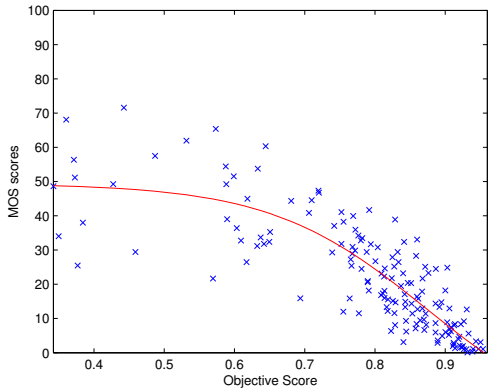


Figure 1.1: Figure showing a scatter plot between MOS from the VQEG dataset and an FR VQA algorithm's scores. A non-linear correlation is evident. Figure also shows a best-fit-line through the scatter obtained using the logistic function proposed in [297].

1.1.2 A Brief Foray into the Human Visual System

You are currently staring at these words on a sheet of paper. Due to acquired fluency in English, it takes you a fraction of a second to view, process, understand and proceed along this page. But it is not language alone that guides you along. The human visual system (HVS) which processes all of the information incident upon the eye and renders it into a form recognizable by higher areas of the human brain for cognitive processes to occur has been one of the most actively researched areas of neuroscience.

The first stage of visual processing in the human are the eyes. This spherical mass is home to different kinds of photoreceptors - receptors that produce a response when incident with photons. The response of these receptors is fed through the retinal ganglion cells and then to the Lateral Geniculate

Neucleus (LGN) which resides in the thalamus. The LGN is analogous to an 'active' switch - receiving and processing both feed-forward and feedback information. LGN responses are passed on to area V1 of the primary visual cortex (situated at the back of your head) which then connects to area V2, V4 as well as area V5/Middle-temporal (MT) and other higher areas in the brain. This kind of hierarchical structure is common in neural processing.

Each of the above described units is an interesting area of study, however we shall not pursue them in detail here. The interested reader is referred to [246] for overviews and descriptions. Here we shall look at these regions of processing using a system-design perspective. The first stage of processing is the human eye. The eye behaves akin to a low-pass filter since light at frequencies above 60 cycles per degree (cpd) are not passed on to the receptors at the back of the eye. Current research indicates that there are two kinds of photoreceptors - rods and cones, based on their response characteristics [303]. Rods are generally in use in low-light conditions while cones are used for vision under well-lit conditions and for color vision. There exist 3 types of cones and depending upon their response characteristics are classified as Long (L), Medium (M) and Short (S) wavelength cones. Another very important characteristic of the eye is the fact that not every region in the visual field is perceived with the same amount of acuity. For example, stare at any one word in this sentence and then try (without moving your eye) to read the beginning of this paragraph. You will notice that even though the word that you are staring at is extremely clear, as you move away from the word under focus,

you start loosing resolution. This is referred to as foveation. If you haven't thought about this before, it may come as a surprise, since the world seems sharp in daily life. This is because the eye performs an efficient engineering solution (given the contraints). The HVS is designed such that the when viewing at a scene, the eye makes rapid movements called saccades interleaved with fixations. Fixations, as the name suggests, refers to the process of looking at a particular location for an extended period of time. Little to no information is gathered during a saccade and most information is gathered during a fixation. Using this strategy of eye movements where the region of maximum visual acuity (fovea) is placed at one location for a short period of time, and then moved to another, the HVS constructs a 'high resolution' map of the scene. Foveation driven video coding is an active area of research [319].

QA systems which seek to emulate the HVS generally model the first stage of processing using a point-spread-function (PSF) to mimic the low-pass response of the human eye. The responses from the receptors in the eye are fed to the retinal ganglion cells. These are generally modeled using center-surround filters, since ganglion cells have been shown to possess on-center off-surround structure [246]. Similar models are used for the LGN. The next stage of the HVS is area V1. The neurons in V1 have been shown to be sensitive to direction, orientation, scale and so on. A multi-scale, multi-orientation decomposition is generally used to mimic this. Better models for V1 involve using multi-scale Gabor filterbanks [247]. The area V5/MT is responsible for processing motion information. Motion estimates are of great importance

for the human since they are used for depth perception, judging velocities of oncoming objects and so on. The engineering equivalent of this region is estimating optical flow [20] from frames in a video. A coarser approximation is block-based motion estimation [226].

In the HVS, the responses from MT/V5 are further sent to higher levels of the brain for processing. We do not discuss them here. The interested reader is referred to [246] for details. Even though the algorithms that are proposed in this dissertation do not seek to explicitly model the HVS,[2] the described algorithms have some relationships with the HVS which we shall explore.

1.2 Contributions

This dissertation is divided into three major section. The first section deals with blind/no-reference image quality assessment, where a revolutionary NR IQA algorithm that is distortion-agnostic is proposed (Chapter 3). The second section extends the proposed framework for NR IQA and demonstrates a novel application – blind distortion-aware perceptually optimized image repair (Chapter 4). We move from images to videos in Chapter 5, where a large scale human study to assess the quality of videos when viewed on mobile devices is described. This timely database and human opinion analysis will guide the development of behavior-aware video quality assessment algorithms. The content of each of these chapters is summarized below.

[2]Since our limited understanding of the HVS makes these HVS-models poor imitations and hence reduce QA algorithm performance

1.2.1 No-reference Image Quality Assessment

Our approach to blind image quality assessment (IQA) is based on the hypothesis that natural scenes possess certain statistical properties which are altered in the presence of distortion, rendering them *un-natural*; and that by characterizing this un-naturalness using scene statistics one can identify the distortion afflicting the image and perform no-reference (NR) IQA. Based on this theory, we propose an (NR)/blind algorithm - the Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index - that assesses the quality of a distorted image without need for a reference image. DIIVINE is based on a 2-stage framework involving distortion identification followed by distortion-specific quality assessment. DIIVINE is capable of assessing the quality of a distorted image across multiple distortion categories, as against most NR IQA algorithms that are distortion-specific in nature. DIIVINE is based on natural scene statistics which govern the behavior of natural images. In this paper, we detail the principles underlying DIIVINE, the statistical features extracted and their relevance to perception and thoroughly evaluate the algorithm on the popular LIVE IQA database. Further, we compare the performance of DIIVINE against leading full-reference (FR) IQA algorithms and demonstrate that DIIVINE is *statistically superior* to the often used measure of peak signal-to-noise ratio (PSNR) and *statistically equivalent* to the popular structural similarity index (SSIM).

1.2.2 Distortion-aware Perceptually Optimized Blind Image Repair

We define the new idea of image repair as a process of correcting one or more possibly different types of distortions afflicting an image. These distortions could introduce linear or non-linear degradations, compression artifacts, noise etc., or combinations of these. Thus the concept encompasses denoising, deblurring, deblocking, deringing, and any other post-acquisition image improvement processes that address distortions. The problem becomes distortion-blind when the nature of the distortion processes is unknown prior to analyzing the image. Towards solving this problem, we describe a new framework for repairing an image that has undergone an unknown set of distortions, based on identifying the distortion(s) present in the image (if any) and applying possibly multiple distortion-specific image repair algorithms. Our philosophy is based on the principle that the task of general purpose image repair is one of agglomeration, i.e., the algorithm should embody multiple high-performing distortion-specific repair modules such that seamless general purpose image repair is achieved. Our proposed framework – the GEneral-purpose No-reference Image Improver (GENII) – is blind to distortion type as well as to distortion parameters, and only requires as input the distorted image to be repaired. GENII is modular and easily extensible to image repair problems beyond those considered here. GENII operates by using natural scene statistic models to identify distortion, to perceptually optimize the distortion parameter(s), to assess the quality of the intermediate repaired images, and

to perceptually optimize the repair processes. We explain the general purpose image repair framework and a realization of this framework, dubbed GENII-1. This implementation assumes that the image has been affected by only a single unknown distortion from among four possibilities. We evaluate the performance of GENII-1 on 4000 distorted images, and demonstrate that it delivers substantial improvements in both quantitative and qualitative visual quality.

1.2.3 Video Quality assessment on Mobile Devices

We shall introduce a new resource that models video distortions in heavily-trafficked wireless networks and that contains measurements of human subjective impressions of the quality of videos. The new LIVE Mobile Video Quality Assessment (VQA) database consists of 200 distorted videos created from 10 RAW HD reference videos, obtained using a RED ONE digital cinematographic camera. While the LIVE Mobile VQA database includes distortions that have been previously studied such as compression and wireless packet-loss, it also incorporates dynamically varying distortions that change as a function of time, such as frame-freezes and temporally varying compression rates. The subjective study portion of the database includes both the differential mean opinion scores (DMOS) computed from the ratings that the subjects provided at the end of each video clip, as well as the continuous temporal scores that the subjects recorded as they viewed the video. The study involved over 50 subjects and resulted in 5,300 summary subjective scores and

time-sampled subjective traces of quality. We also study a variety of models of temporal pooling that may reflect strategies that the subjects used to make the final decision on video quality. Further, we compare the quality ratings obtained from the tablet and the mobile phone studies in order to study the impact of these different display modes on quality. We also evaluate several objective image and video quality assessment (IQA/VQA) algorithms with regards to their efficacy in predicting visual quality. A detailed correlation analysis and statistical hypothesis testing is carried out.

Before we describe the above contributions in detail, a short literature review of the relevant fields is undertaken.

Chapter 2

Literature Review

This chapter performs a brief overview of previous work in the topics to be described in the rest of this dissertation. This chapter is by no means comprehensive and only summarizes relevant literature, while pointing the interested reader to more comprehensive reviews.

2.1 No-reference Image Quality Assessment

Most present-day NR IQA algorithms assume that the distorting medium is known - for example, compression, loss induced due to noisy channel etc. Based on this assumption, distortions specific to the medium are modeled and quality is assessed. By far the most popular distorting medium is compression which implies that blockiness and bluriness should be evaluated. In the following, we study blind QA algorithms that target three common distortion categories: JPEG compression, JPEG2000 compression, and blur. We also survey blind QA algorithms that operate holistically.

2.1.1 Distortion-specific IQA Algorithms

2.1.1.1 JPEG IQA

The general approach to NR JPEG IQA is to measure edge strength at block boundaries and relate this strength and possibly some measure of image activity to perceived quality. JPEG NR IQA algorithms include those that use a hermite transform based approach to model blurred edges [167], those that estimate first-order differences and activity in an image [316], those that utilize an importance map weighting of spatial blocking scores [19], those that use a threshold-based approach on computed gradients [52] and those that compute block strengths in the Fourier domain [279]. Each of these approaches measures a subset of blocking, blur and activity and computes perceptual quality, either using a training set, or by combining features in an intelligent fashion.

2.1.1.2 JPEG2000 IQA

For JPEG2000 ringing artifacts in an image are generally modeled by measuring edge-spread using an edge-detection based approach and this edge spread is related to quality [204], [286], [163]. Other approaches include those that compute simple features in the spatial domain [242], or those that utilize natural scene statistics [262]. In [262], the authors exploit the dependency between a wavelet coefficient and its neighbors, and the fact that the presence of distortion will alter these dependencies. The dependencies are captured using a threshold + offset approach, where the parameters are estimated using a training set.

2.1.1.3 Sharpness/Blur IQA

Blur IQA algorithms model edge spreads and relate these spreads to perceived quality, similar to the approach followed by NR JPEG2000 IQA algorithms. Edge strengths are quantified using a variety of techniques, including block kurtosis of DCT coefficients [36], iterative thresholding of a gradient image [296], and measuring the probability of blur detection [188] or model the just-noticeable-blur [81] in an image. Researchers have also explored the use of saliency models for NR blur IQA [236]. A noise-robust blur measure was also proposed in [344] that utilizes a gradient-based approach coupled with the singular value decomposition.

It should be clear to the reader that each of these distortion specific NR IQA algorithms attempt to model indicators of quality for the distortion in question, and hence are unsuitable for use in a general-purpose (distortion-agnostic) scenario.

2.1.2 Holistic IQA algorithms

Li proposed a series of heuristic measures to characterize image quality based on three quantities - edge sharpness, random noise level (impulse/additive white Gaussian noise) and structural noise [143]. Edge sharpness is measured using an edge-detection approach, while the random noise level is measured using a local smoothness approach (impulse noise) and PDE-based model (Gaussian noise). Structural noise as defined by Li relates to blocking and ringing from compression techniques such as JPEG and JPEG2000. Unfortunately,

the author does not analyze the performance of the proposed measures, nor propose a technique to combine the measures to produce a general-purpose quality assessment algorithm.

Gabrada and Cristobal proposed an innovative strategy for blind IQA which utilized the Renyi entropy measure [88] along various orientations to measure anisotropy. The proposed approach is attractive since natural images are anisotropic in nature and possesses statistical structure that distortions destroy. They measure mean, standard deviation and range of the Renyi entropy along four pre-defined orientations in the spatial domain and demonstrate their correlation with perceived quality. Unfortunately, a thorough evaluation of the proposed measure is again lacking.

Recently, Saad and Bovik proposed a general-purpose blind quality assessment algorithm that computes four features in the DCT domain: DCT kurtosis, DCT contrast and two anisotropy measures inspired from [88] - maximum and variance of the Renyi entropy along four orientations [235]. Features are extracted over two scales and a Gaussian distribution is used to model the relationship between the DMOS and the extracted features. The measure was shown to perform well in terms of correlation with human perception across distortion categories.

Blind/NR *video* quality assessment (VQA) is an important problem that has followed a similar trajectory. Some authors have proposed techniques which measure blockiness, blur, corner outliers and noise separately, and use a

Minkowski sum to pool the measures of quality together [37, 79]. In both these approaches, distortion-specific indicators of quality are computed and pooled using a variety of pre-fixed thresholds and training, as against our approach that uses concepts from NSS to produce a modular and easily extensible approach that can be modified to include other distortions than those discussed here. We anticipate that the approach taken here could be eventually extended to video to achieve good results.

2.2 Image Repair

The diverse subfields of image repair that we address have been well explored and broadly surveyed and hence we refrain from a thorough review of these techniques. Instead, we summarize some key algorithms and where appropriate point the reader to other literature in the field. We (very briefly) summarize relevant research on the following subclasses of image repair : (1) deringing, (2) deblocking, (3) denoising and (4) deblurring/deconvolution. While we are unaware of algorithms that tackle all the four distortions considered here, some algorithms tackle more than one of these distortions and this is noted in the summary below.

Deringing Existing approaches include iterative projection on to convex sets (POCS) [144], total variation [134, 189, 190], anisotropy [23], bilateral filtering and its variants [75, 131, 193, 285] and quadtree decompositions [55, 56, 340] .

Deblocking Prominent deblocking algorithms include those that use a field of experts model for natural images [276, 277], those operating in the DCT domain [342], those that use local smoothing filters [27], and block processing [146, 341] and those based on POCS [147]. A shape adaptive DCT algorithm for denoising has also been shown to perform well at deblocking images [83].

Denoising Algorithms for denoising include subband methods [137, 222–224, 271], those that use sparse coding [76] and those based on collaborative filtering and local shape adaptation in the DCT domain [61, 83]. Reviews and analysis may be found in [29, 49].

Deblurring Deconvolution algorithms include approaches based on collaborative Wiener filtering [62], statistics of natural images in the gradient domain [140, 145], color statistics of natural images [119], space-variant Gaussian scale mixture (GSM) statistical modeling of wavelet coefficients [97, 98] and those that tackle spatially varying blur [16]. The approaches in [97, 98, 119] (among others) are capable of performing both denoising and deblurring.

2.3 Video Quality Assessment on Mobile Devices

Several researchers have conducted subjective video quality studies with various aims [182, 255, 297, 298, 300]. Significant effort has also been applied to designing objective algorithms that are capable of predicting visual quality with high correlation against subjective perception [219, 252, 313, 324]. Previ-

ous subjective studies on VQA have been performed on large format displays such as CRT/LCD monitors, while typically distorted videos have included compressed videos (H.264/MPEG), videos transmitted over wireless/IP channels [255],[182] and jittered and delayed videos [99, 112]. While video quality on mobile devices has not been extensively researched, there have been a few studies on the quality assessment of videos on mobile devices.

Eichhorn and Ni performed a human study to evaluate the quality of H.264 scalable video codec (SVC) encoded video streams at QVGA and QQVGA resolutions on a 2.5-inch screen [74]. Each of the six 8-second clips were encoded at two spatial resolutions using 3 temporal layers and 4 quality layers. Thirty subjects rated the visual quality of the videos yielding a differential mean opinion (DMOS) score for each of the videos in the database. Based on the DMOS obtained, the authors analyzed the effect of reduced spatial resolution as well as reduced temporal sampling and quality. While the analysis presented is interesting, the low-resolution of the videos (QVGA/QQVGA) relative to those displayed by current mobile devices, the fact that some of the videos in the database were un-natural (eg., animations) and the unavailability of the database limit its current utility.

Knoche and colleagues evaluated image resolution requirements for MobileTV by conducting a large-scale human study where over 120 subjects participated (although each video only received 32 ratings) [133]. The subjects were asked to rate the quality of videos which had gracefully decreasing encoding bit-rates (using Microsoft Windows Video V8 codec) and varying res-

olutions on a display of resolution 240×320. The results presented are quite valuable, especially since the authors also varied audio quality in the study. However, from an algorithm design-perspective, the lack of pristine reference videos as well as the manner in which some of the videos were artificially modified (eg., feeds from News which included text scrolls, picture-in-picture etc.), coupled with its unavailability again limits the usefulness of the database.

Jumisko-Pyykko and Hakkinen performed a subjective study where reference clips from video tapes were converted to digital video, then compressed using a variety of video codecs (H.263, H.264 etc.) [120]. The authors evaluated video-only as well as audio-video quality on the Nokia 6600 and the S-E P800. As with other studies of this nature, the very low frame-rates and bit-rates relative to current technology and the lack of public availability reduce the currency of the work.

Ries et al. evaluated the quality of five reference videos of 10-seconds each when compressed at varying frame-rates and bit-rates using the H.264/AVC baseline encoder [227]. The authors also detailed an algorithm that would evaluate the quality of these videos so that the objective scores produced would correlate well with the obtained human opinion scores. All of the limitations of the above databases apply to this one as well. Other studies on mobile devices include an investigation on context and its effect on quality [121], and a study of the effect of extremely low bit-rates on perceived quality [328].

Having summarized the previous work in the relevant areas, we now

move on to the crux of this dissertation, and in a series of chapters, describe the contributions.

Chapter 3

Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality

We have developed a computational theory for NR IQA based on the statistics of natural images[1] [93, 202, 269, 270]. Natural un-distorted images possess certain statistical properties that hold across different image contents. For example, it is well known that the power spectrum of natural scenes fall-off as (approximately) $1/f^{\gamma}$, where f is frequency [93]. Natural scene statistic (NSS) models seek to capture those statistical properties of natural scenes that hold across different contents. Our approach to NR IQA is based on the hypothesis that, the presence of distortions in natural images alters the natural statistical properties of images, thereby rendering them (and consequently their statistics) *unnatural* [261]. The goal of an NR IQA algorithm based on NSS is to capture this 'unnatural-ness' in the distorted image and relate it to perceived quality. In the past, such NSS-based QA algorithms have been successfully deployed for FR IQA [261, 263], for RR IQA [142, 321] and to a small extent, for NR IQA [262]. We explore such an NSS-based approach for NR IQA.

[1]By natural we mean any image that can be obtained from a camera - these include pictures of man-made objects as well as forest/natural environments.

Our NR IQA model utilizes a 2-stage framework for blind IQA that we introduced in [178]. In this framework, scene statistics extracted from a distorted natural image are used to first explicitly classify the distorted image into one of n distortions (distortion identification - stage 1). Then, the same set of statistics are used to evaluate the distortion-specific quality (distortion-specific QA - stage 2) of the image. A combination of the two stages leads to a quality score for the image which, as we shall soon demonstrate, correlates quite well with human perception and is competitive with leading FR IQA algorithms. The proposed approach we call Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE). The name is appropriate as the algorithm resulting from the modeling framework succeeds at 'divining' image quality without any reference information or the benefit of distortion models.

The DIIVINE approach to NR IQA is a full-fledged realization of the preliminary framework that we had proposed in [178]. In [178], we had primarily proposed the 2-stage framework and demonstrated simple implementations of the framework as examples. Apart from the fact that the DIIVINE approach performs much better than those realizations, the main contribution of this work is the series of statistical features that we extract, which go beyond the simple marginal descriptions that the previous primary realizations extracted.

Before proceeding, we state some salient aspects of DIIVINE. Present-day NR IQA algorithms are distortion-specific, i.e., the algorithm is capable of assessing the quality of images distorted by a particular distortion type. For

example, the algorithm in [167] is for JPEG compressed images, that in [242] is for JPEG2000 compressed images and that in [36] is for blur. DIIVINE, however, is not bound by the distortion-type affecting the image since we do not seek distortion-specific indicators of quality (such as edge strength at block boundaries) but provide a modular strategy that adapts itself to the distortion in question. Indeed, our framework is ostensibly distortion-agnostic.

Further, since we do not use distortion-specific models, DIIVINE can easily be extended to handle distortions beyond those considered here. Finally, by performing a thorough analysis of our algorithm we demonstrate that DIIVINE is competitive with present day NR and FR IQA algorithms *across* commonly encountered distortions. In fact, we shall demonstrate that DIIVINE is not only statistically superior to the full-reference peak signal-to-noise-ratio (PSNR) measure of quality, but is also statistically indistinguishable from a popular full-reference measure - the structural similarity index (SSIM) [311].

3.1 Scene Statistics of Distorted Images

The DIIVINE approach for NR IQA proceeds as follows. The distorted image is first decomposed using a scale-space-orientation decomposition (loosely, a wavelet transform) to form oriented band-pass responses. The obtained subband coefficients are then utilized to extract a series of statistical features. These statistical features are stacked to form a vector which is a statistical description of the distortion in the image. Our goal is to utilize these

feature vectors across images to perform two tasks in sequence: (1) Identify the probability that the image is afflicted by one of the multiple distortion categories, then (2) Map the feature vector onto a quality score for each distortion category, i.e., build a regression model for each distortion category to map the features onto quality, conditioned on the fact that the image is impaired by that particular distortion category (i.e., distortion-specific QA). The probabilistic distortion identification estimate is then combined with the distortion-specific quality score to produce a final quality value for the image. The method described here is illustrated in Fig. 3.1 and is labeled as the Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index.

3.1.1 Statistical Model for Wavelet Coefficients

In the DIIVINE framework, a set of neighboring wavelet coefficients are modeled using the Gaussian Scale Mixture (GSM) model [302]. An N dimensional random vector Y is a GSM if $Y \equiv z \cdot U$ where $\equiv$ denotes equality in probability distribution, U is a zero-mean Gaussian random vector with covariance C_U, and z is a scalar random variable called a mixing multiplier. The density of Y is then given by:

$$p_Y(y) = \int \frac{1}{(2\pi)^{N/2}|z^2 C_U^{1/2}|} exp(\frac{-Y^T C_U^{-1} Y}{z^2}) p_Z(z) dz$$

The GSM model has been used to model the marginal and joint statistics of the wavelet coefficients of natural images [142, 302], where the vector

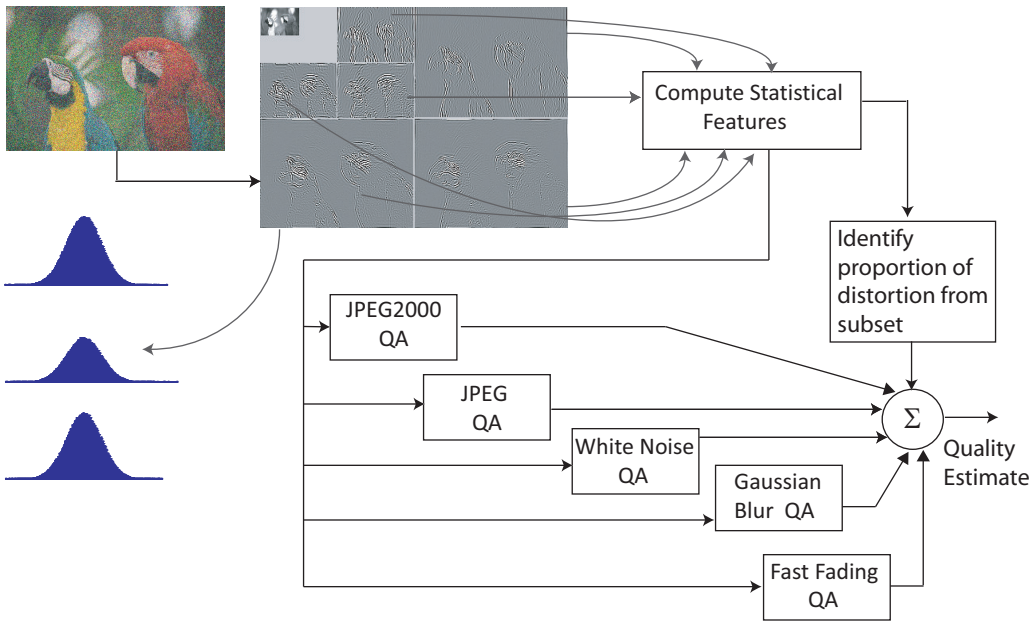


Figure 3.1: The proposed Distortion identification-based Image Verity and INtegrity Evaluation (DIIVINE) index consists of two stages: probabilistic distortion identification followed by distortion-specific quality assessment as illustrated here.

Y is formed by clustering a set of neighboring wavelet coefficients within a subband, or across neighboring subbands in scale and orientation.

Next, we shall describe the statistical features that we extract from the distorted image and motivate their choice. In order to illustrate how each of these features behaves in natural and distorted images, we shall use the natural un-distorted reference images shown in Fig. 3.2 and their distorted counterparts, a subset of which are shown in Fig. 3.3. The distortions in Fig. 3.3 are exactly the same as the ones that we consider in this paper and that span the set of distortions in the LIVE Image Quality Assessment Database

(a)

(b)

(c)

(d)

Figure 3.2: (a)-(d) The images used to demonstrate features derived under NSS models.

[264] - JPEG and JPEG2000 (JP2k) compression, additive white noise (WN), Gaussian blur (blur) and a Rayleigh fading channel labeled fast fading (FF).

3.1.2 Extracting Scene Statistics

In order to extract statistics from distorted images we utilize the steerable pyramid decomposition [268]. The steerable pyramid is an overcomplete wavelet transform that allows for increased orientation selectivity. The choice of the wavelet transform was motivated by the fact that the scale-space-orientation decomposition that the wavelet transform performs mirrors models

(a)

(b)

(c)

(d)

(e)

Figure 3.3: A subset of the distorted versions of images in Fig. 3.2. (a)-(e) correspond to the following distortions - (a) JP2k compression, (b) JPEG compression, (c) white noise, (d) Gaussian blur and (e) fast fading distortion.

of spatial decomposition that occurs in area V1 of the primary visual cortex [203, 246]. The steerable pyramid has been previously used for FR IQA [261] as well as RR IQA [142] with success. Note that we do not use the complex version of the steerable pyramid as in [240], but that used in [261].

Given an image whose quality is to be assessed, the first step is to perform a wavelet decomposition using a steerable pyramid over 2 scales and 6 orientations. We have found that an increased degree of orientation selectivity is beneficial for the purpose of QA - more so than selectivity over more than 2 scales. The choice of steerable filters was also motivated by its increased orientation selectivity. Our experiments have indicated that increasing the number of scales beyond 2 does not improve performance. The resulting decomposition results in 12 subbands across orientations and scales labeled s_α^θ, where $\alpha \in \{1, 2\}$ and $\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$.

The next step is to perform the perceptually significant process of divisive normalization [301]. Divisive normalization or contrast-gain-control was proposed in the psychovisual literature in order to account for the non-linear behavior of certain cortical neurons. Such normalization accounts for interactions between neighboring neurons and governs the response of a neuron based on the responses of a pool of neurons surrounding it [301]. Divisive normalization also reduces the statistical dependencies between subbands thereby de-coupling subband responses to a certain degree [301, 302]. Further, divisive normalization models partially account for contrast masking [246] - an essential ingredient in QA algorithm design. Divisive normalization has been

explicitly used for RR IQA in the past [142]. FR IQA techniques such as the visual information fidelity index (VIF) [261] and the structural similarity index (SSIM) [311] also utilize divisive normalization, albeit in an implicit manner [250]. Finally, the successful MOVIE index, a recently proposed FR VQA algorithm [252] also utilizes such a technique (drawing inspiration from the Teo and Heeger model [283]). Here, divisive normalization is implemented as described in [142].

Specifically, given a subband coefficient y, our goal is to compute a normalization parameter p, based on responses from neighboring subbands in order to finally compute $\hat{y} = y/p$. To estimate p we utilize the previously defined local statistical model for natural images - the Gaussian scale mixture (GSM) model [302]. In our implementation, for a center coefficient y_c at each subband we define a divisive normalization transform (DNT) neighborhood vector Y that contains 15 coefficients, including 9 from the same subband (3×3 neighborhood around y_c), 1 from the parent band, and 5 from the same spatial location in the neighboring bands at the same scale. Given this vector Y, the normalization coefficient p is computed as $p = \sqrt{Y^T C_U^{-1} Y / N}$. This computation is undertaken at each coefficient in each subband to produce a divisively-normalized set of subbands - d_α^θ, where $\alpha \in \{1, 2\}$ and $\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$. The interested reader is referred to [142] for details on the divisive normalization procedure.

In order to visualize how divisive normalization affects the statistics of the subband coefficients, Fig. 3.4 plots a histogram of coefficents from s_1^θ and

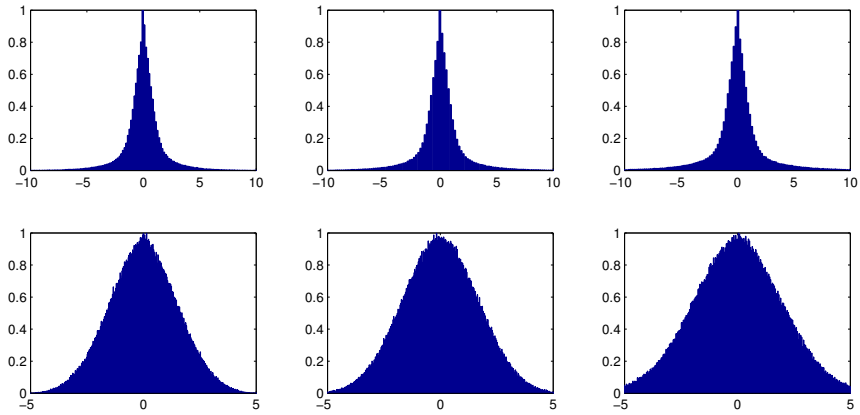


Figure 3.4: Figure demonstrating the effect of divisive normalization on the subband statistics of image in Fig. 3.2(a). The first row shows the histogram of subband coefficient distributions before divisive normalization, while the second row is the distribution after normalization. Divisive normalization makes the subband statistics of *natural images* more Gaussian-like, as compared to the Laplacian nature of the pre-normalized subband coefficients.

d_1^θ, where $\theta \in \{0°, 30°, 60°\}$. The normalization makes the subband statistics more Gaussian-like for natural images.

In order to demonstrate that subband statistics are affected by each distortion in a particular fashion, Fig. 3.5 plots the coefficient distributions from $d_1^{0°}$ of the image in Fig. 3.2(c) for each distortion considered here. It should be clear that each distortion affects the statistics in a characteristic way which is essentially independent of the content (e.g., WN always increases the variance of subband coefficients).

Given that each distortion affects subband statistics characteristically, the goal is to compute marginal and joint statistics across subbands in order

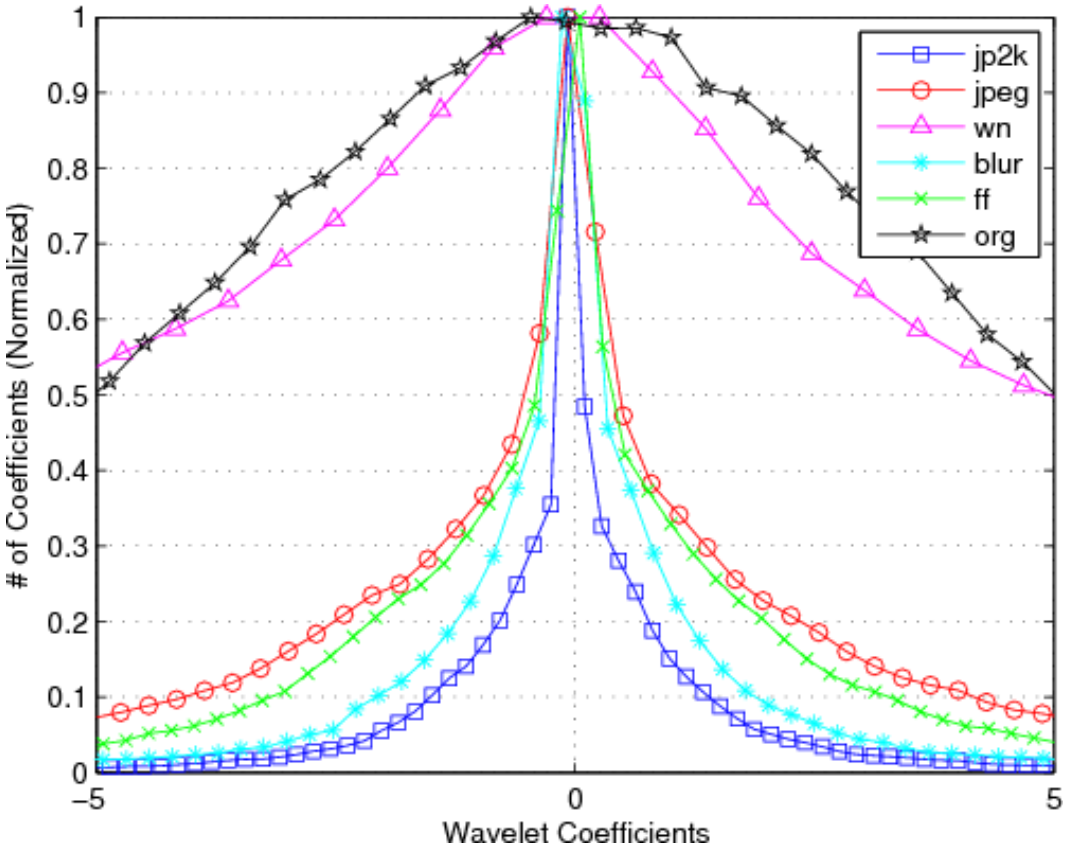


Figure 3.5: Subband statistics from $d_1^{0^\circ}$ of the image in Fig. 3.2 (c) for different distortions. Notice how each distortion affects the statistics in a characteristic way.

to extract features that are relevant to the perceived quality of the image.

3.1.2.1 Scale and orientation selective statistics (f_1-f_{24})

Subband coefficients from each of the 12 subbands are parametrized using a generalized Gaussian distribution (GGD). The GGD is:

$$f_X(x;\mu,\sigma^2,\gamma) = ae^{-[b|x-\mu|]^\gamma} \qquad x \in \Re$$

where, μ, σ^2 and γ are the mean, variance and shape-parameter of the distribution and

$$a = \frac{b\gamma}{2\Gamma(1/\gamma)}$$

$$b = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}$$

and $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \quad x > 0$$

The shape parameter γ controls the 'shape' of the distribution. For example, $\gamma = 2$ yields a Gaussian distribution and $\gamma = 1$ yields a Laplacian distribution. The parameters of the distribution (μ, σ^2 and γ) are estimated using the method proposed in [258]. GGD has also been used before to model the subband statistics of natural images in RR IQA [321]. Since wavelet subband responses are zero mean, we have to estimate σ^2 and γ for each subband leading to a total of 24 features. f_1-f_{12} correspond to σ^2 across subbands and f_{13}-f_{24} correspond to γ across subbands.

At this juncture it may be prudent to explain the choice of the GGD. The divisive normalization procedure tends to produce coefficients distributed in a Gaussian manner for *natural* images. In the presence of distortion however, this Gaussianity at the output of the normalization procedure is not guaranteed. For example, from Fig. 3.5, it should be clear that distortions such as JPEG, JP2k, Blur and FF lead to highly kurtotic (non-Gaussian) distributions even after the divisive normalization procedure. Since the shape parameter of the GGD will capture this non-Gaussian nature, the GGD fit is utilized here as against a simple Gaussian fit. We note that a similar procedure was used for RR IQA in [142].

In order to demonstrate how these subband features affect quality, Fig. 3.6 shows a plot of $log_e(\sigma^2)$ $vs.$ γ for one of the subbands for each of the reference images in Fig. 3.2 and their associated distorted versions.

We have previously shown that these simple marginal statistics when used in a simple, preliminary blind IQA algorithm - the Blind Image Quality Index (BIQI) [178] - do a good job of identifying the distortion afflicting the image and predicting the perceived quality of an image [177, 178]. Here we full develop the 2-stage NSS-based IQA concept introduced in [178], by deploying a much richer set of NSS-based features that capture the dependencies between subband coefficients over scales and orientations, as well as utilizing the perceptually relevant divisive normalization procedure.

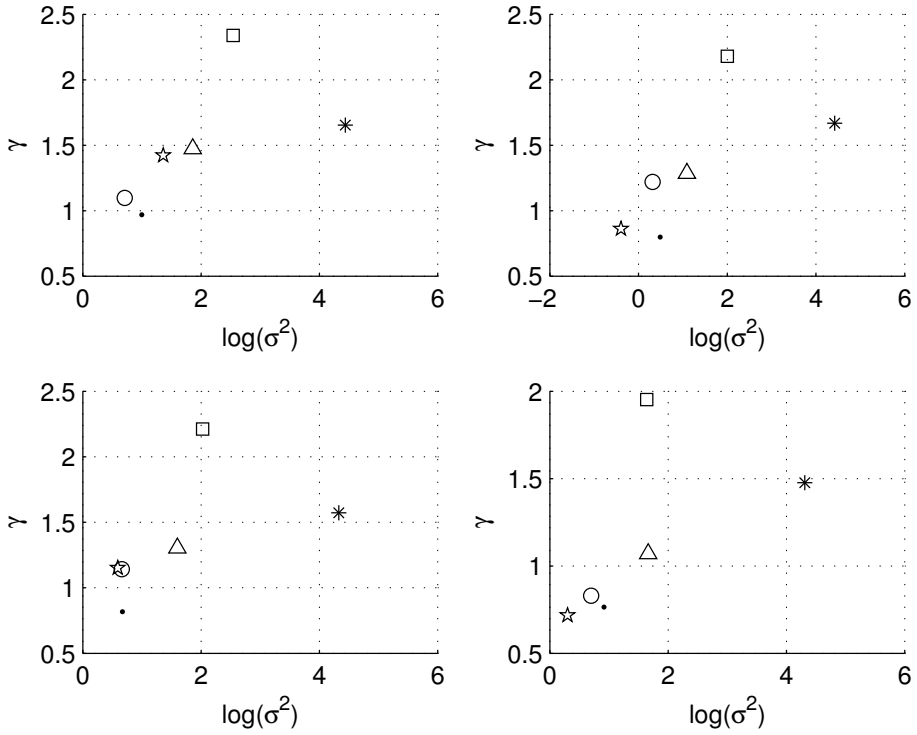


Figure 3.6: Plot of $log_e(\sigma^2)$ *vs.* γ for $d_1^{120^\circ}$ for each of the images considered in Fig. 3.2 and their associated distorted versions. Notice how each distortion seems to cluster in a particular region, immaterial of the image content. Images (a)-(d): left to right, top to bottom. Reference image ($\square$), JPEG ($\circ$), JPEG2K (Δ), WN ($*$), blur ($\cdot$) and ff ($\star$).

3.1.2.2 Orientation selective statistics (f_{25}-f_{31})

Images are naturally multiscale. Further, there exists a relationship between subbands at the same orientation and across different scales. Distortions in an image will affect these across-scale statistics. For example, in Fig. 3.7, we plot a histogram of coefficients from $d_1^{0^\circ}$ and $d_2^{0^\circ}$ for the image in Fig. 3.2(c) and its various distorted versions. In order to plot these distributions in 1-D, these subbands were stacked together to form a large vector, whose histogram we plot. Notice the difference in distributions of these across-scale coefficients for natural and distorted images.

In order to capture the variation seen in Fig. 3.7, we again utilize a GGD fit. The 1-D GGD is now fit to the coefficients obtained by stacking together coefficients from subbands at the same orientation but at different scales. Specifically, 6 GGD fits corresponding to each one of $\{d_1^\theta, d_2^\theta\}$, $\theta \in \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ\}$ are computed. Again, these fits are zero-mean and we compute two parameters - σ^2 and γ. In our experiments, σ^2 does not add any information about the perceived quality and hence we use only the computed γ's as features. Further, we also compute a GGD fit when all of the subbands are stacked together (i.e., $\{d_\alpha^\theta\}, \forall \alpha, \theta$) and use the γ parameter again as our feature. Thus, $f_{25} - f_{30}$ correspond to γ from the statistics across scales over different orientations, while f_{31} corresponds to γ from the statistics across subbands. In Fig. 3.8 we plot these computed γ values for each of the images in Fig. 3.2 and their associated distorted versions.

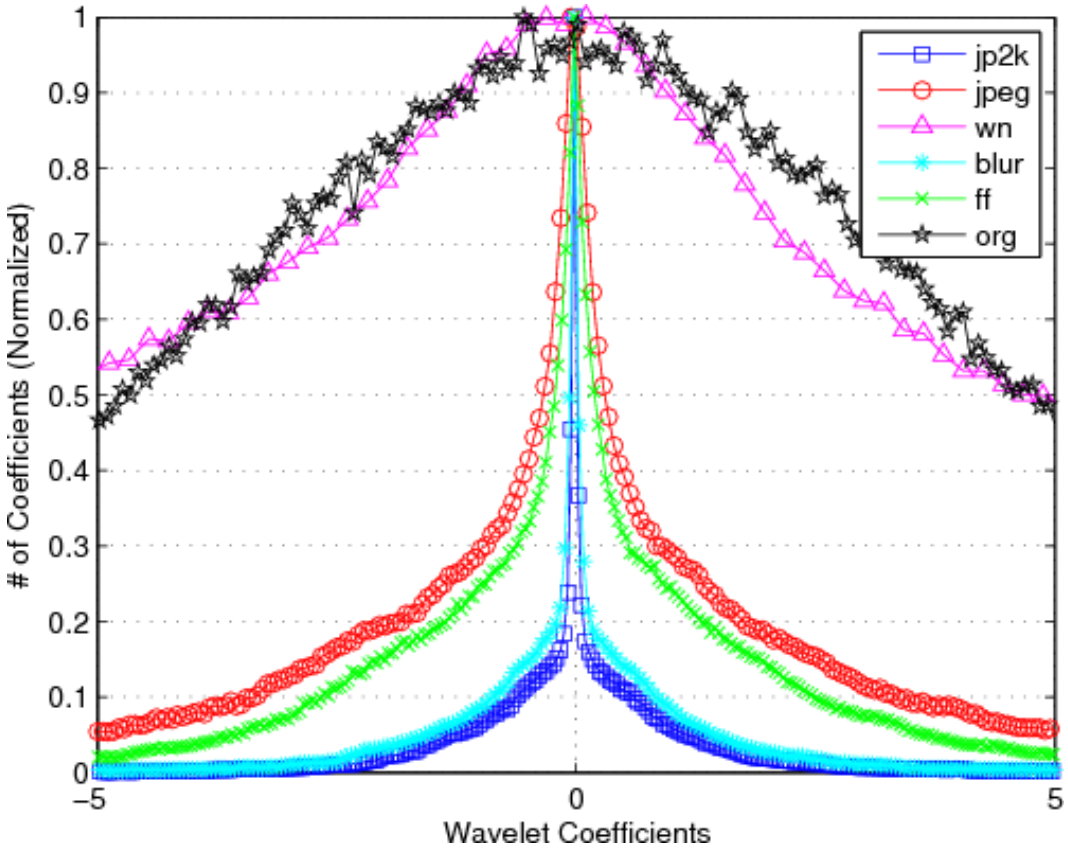


Figure 3.7: Histogram (normalized) of coefficients from $d_1^{0^\circ}$ and $d_2^{0^\circ}$ for the image in fig. 3.2(c) and its various distorted versions. Notice the difference in distributions of these across-scale coefficients for natural and distorted images.

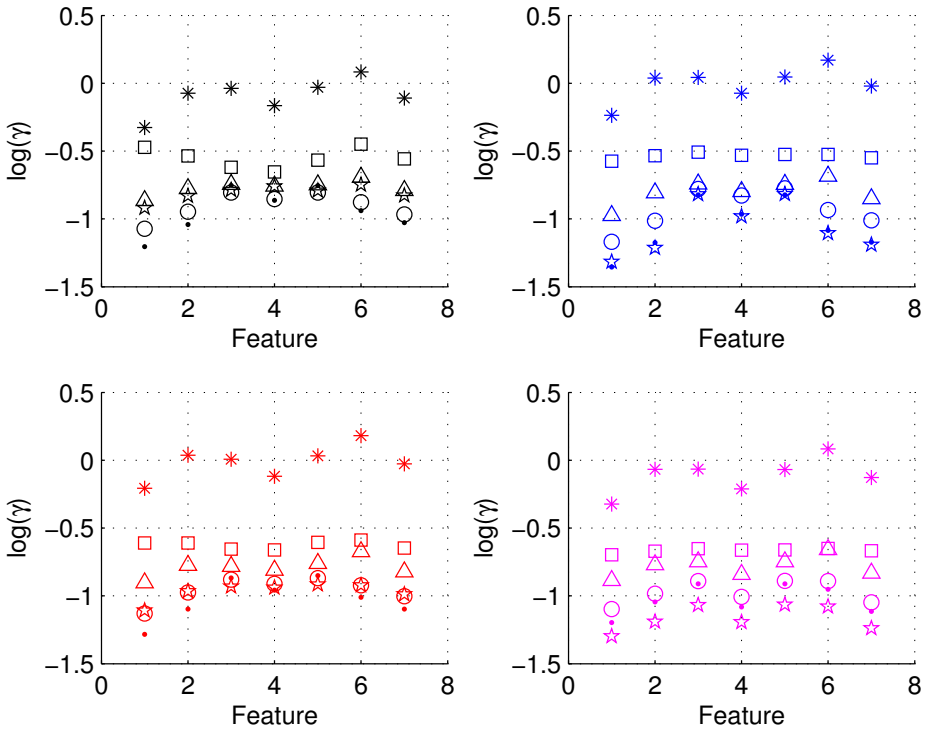


Figure 3.8: Orientation Selective Statistics (γ) for reference and distorted images. Images (a)-(d): left to right, top to bottom. Reference image ($\square$), JPEG ($\circ$), JPEG2K (Δ), WN ($*$), blur ($\cdot$) and ff ($\star$). $1-7$ on the x-axis correspond to f_{25}-f_{31}.

3.1.2.3 Correlations across scales (f_{32}-f_{43})

One of the primary stages in human visual processing is filtering of the visual stimulus by the retinal ganglion cells [246]. These cells have center-surround-difference properties and have spatial responses that resemble difference of Gaussians (DoG) functions [208, 246]. The responses of these cells serve a variety of likely purposes including dynamic range compression, coding and enhancement of features such as edges [208, 246]. Image compression algorithms such as EZT and SPIHT [237, 257] offer evidence of correlations across scales as well. Statistics of edges have been used for blur quality assessment [53]. Given that edges are important, it is reasonable to suppose that there exist elegant statistical properties between high-pass responses of natural images and their band-pass counterparts. Indeed, in our experiments, we found that such a relationship exists for natural images and this relationship is affected by the presence of distortion. We model high-pass band-pass correlations in order to capture these dependencies.

Each bandpass (BP) subband is compared with the high-pass (HP) residual band (obtained from the steerable pyramid transform) using a windowed structural correlation [311]. Specifically, the BP and HP bands are filtered using a 15×15 Gaussian window with $\sigma = 1.5$ [311]. The structural correlation is then computed as:

$$\rho = \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

where σ_{xy} is the cross-covariance between the windowed regions from the BP

and HP bands, and σ_x^2, σ_y^2 are their windowed variances respectively; C_2 is a stabilizing constant that prevents instabilities from arising when the denominator tends to 0, and its value is the same as that used in [311]. The mean of the correlation map so obtained is used as the correlation feature.

Fig. 3.9 plots the value of the correlation coefficient for each of the 12 subbands and for all images considered in fig. 3.2 and their associated distorted versions. Again, distortion-specific clustering immaterial of content is evident.

Since there are 12 subbands, 12 such correlations are computed, yielding features f_{32}-f_{43}.

3.1.2.4 Spatial correlation (f_{44}-f_{73})

Throughout this discussion we have emphasized the observation that natural images are highly structured and that distortions modify this structure. While we have captured many such modifications in the subband domain, one particular form of scene statistics that remains neglected is the spatial structure of the subbands. Natural images have a correlation structure that, in most places, smoothly varies as function of distance.

In order to capture spatial correlation statistics, we proceed as follows. For each τ, $\tau \in \{1, 2, \ldots, 25\}$, and for each d_1^θ, $\theta \in \{0°, 30°, 60°, 90°, 120°, 150°\}$, we compute the joint empirical distribution between coefficients at (i, j) and $\mathcal{N}_8^\tau(i, j)$, where $\mathcal{N}_8^\tau$ denotes the set of spatial locations at a distance of τ (chessboard distance). The joint distribution attained for a value of τ can be thought

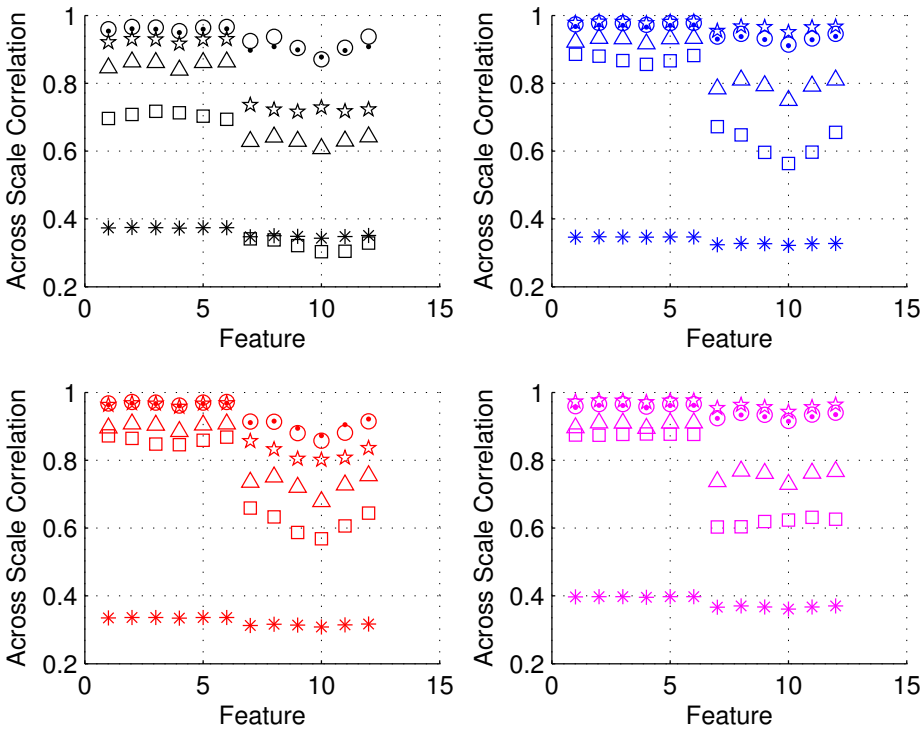


Figure 3.9: Across scale correlation statistics for reference and distorted images. Images (a)-(d): left to right, top to bottom. Reference image ($\square$), JPEG ($\circ$), JPEG2K (Δ), WN ($*$), blur ($\cdot$) and ff ($\star$). $1-12$ on the x-axis correspond to f_{32}-f_{43}.

of as the joint distribution $p_{XY}(x, y)$ between two random variables X and Y. To estimate correlation between these two variables, we compute:

$$\rho(\tau) = \frac{E_{p_{XY}(x,y)}[(X - E_{p_X(x)}[X])^T(Y - E_{p_Y(y)}[Y])]}{\sigma_X \sigma_Y}$$

where $E_{p_X(x)}[X]$ is the expectation of X with respect to the marginal distribution $p_X(x)$ obtained from the computed joint distribution, and similarly for Y and (X, Y). In order to visualize this $\rho(\tau)$ for different distortions, in Fig. 3.10 we plot ρ as a function of τ for the image in Fig. 3.2 (b) and its distorted versions in Fig. 3.3. Notice how the presence of distortion alters the spatial correlations statistics.

Once $\rho(\tau)$ is obtained, we parameterize the obtained curve by fitting it with a 3^{rd} order polynomial, where τ is the distance at which the estimate of ρ is computed. Such a fit is computed for $d_1^\theta, \forall\theta$. The coefficients of the polynomial and the error between the fit and the actual $\rho(\tau)$ form the features - f_{44}-f_{73}.

3.1.2.5 Across orientation statistics (f_{74}-f_{88})

One set of statistics that remains unexplored are statistical correlations that natural images exhibit across orientations. In order to capture the distortion-induced modifications to these statistical correlations across orientations, we compute windowed structural correlation (same as the across scale statistics) between all possible pairs of subbands at the coarsest scale. The set of features is the lowest 5% $[175, 219]$ of the structural correlation values

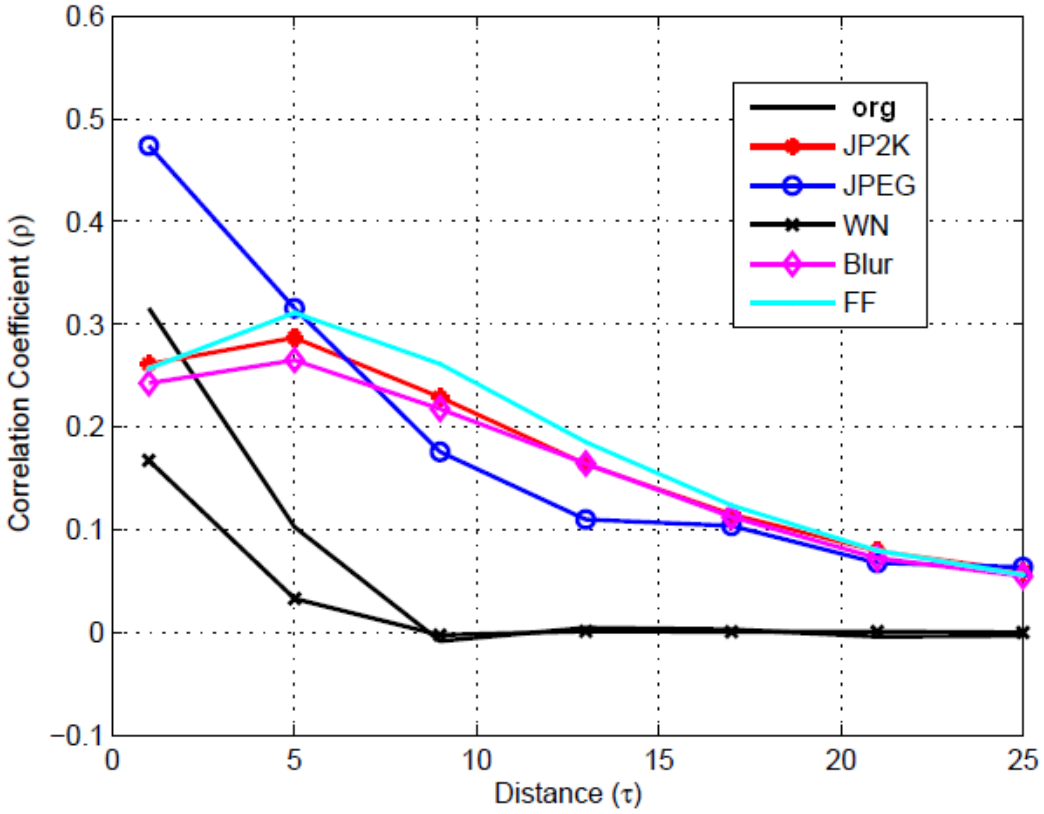


Figure 3.10: Plot of spatial correlation coefficient ($\rho(\tau)$) for various distance τ for one subband of an image, across distortions.

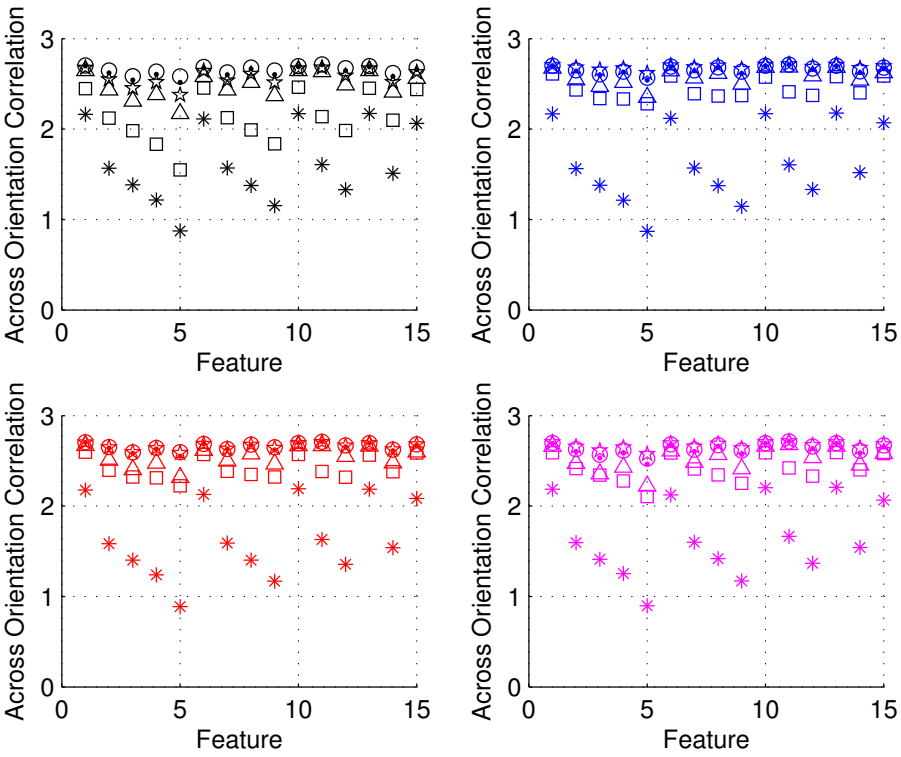


Figure 3.11: Across-orientation statistics for reference and distorted images. Images (a)-(d): left to right, top to bottom. Reference image (□), JPEG (∘), JPEG2K (Δ), WN (∗), blur (·) and ff (⋆). $1-15$ on the x-axis correspond to f_{74}-f_{88}.

so obtained for each pair, leading to a total of ${}^6C_2 = 15$ features - f_{74}-f_{88}. In Fig. 3.11 we plot the value of these across orientation features for each of the images considered in Fig. 3.2 and their associated distorted versions. Notice clustering of distortions independent of content. All of the features described here are listed in Table 3.1 for reference.

Until now, we defined a series of statistical features that we extracted

Feature ID	Feature Description	Computation Procedure
$f_1 - f_{12}$	Variance of subband coefficients	Fitting a generalized Gaussian to subband coefficients
$f_{13} - f_{24}$	Shape parameter of subband coefficients	Fitting a generalized Gaussian to subband coefficients
$f_{25} - f_{31}$	Shape parameter across subband coefficients	Fitting a generalized Gaussian to orientation subband coefficients
$f_{32} - f_{43}$	Correlations across scales	Computing windowed structural correlation between filter responses
$f_{44} - f_{73}$	Spatial correlation across subbands	Fitting a polynomial to the correlation function
$f_{74} - f_{88}$	Across orientation statistics	Computing windowed structural correlation between adjacent orientations at same scale

Table 3.1: Table listing each of the features considered here and the method in which they were computed.

from subband coefficients and we described how each of these statistics are affected in the presence of distortion. However, the relationship to quality for each of these features requires clarification. Hence, in Fig. 3.12 we plot the Spearman's rank ordered correlation coefficient (SROCC) across each of the distorted categories across all distorted images in the LIVE image database. Note that *no training is undertaken here*; the plot is simply to justify the choice of the features as good indicators of quality. As is clear, some features predict perceived quality with greater correlation with human perception than others.

3.2 Distortion-identification based image verity and integrity evaluation

Our 2-stage approach to NR IQA - as realized here in constructing the DIIVINE index - consists of utilizing the features extracted as described above for distortion-identification as well as for distortion-specific quality assessment

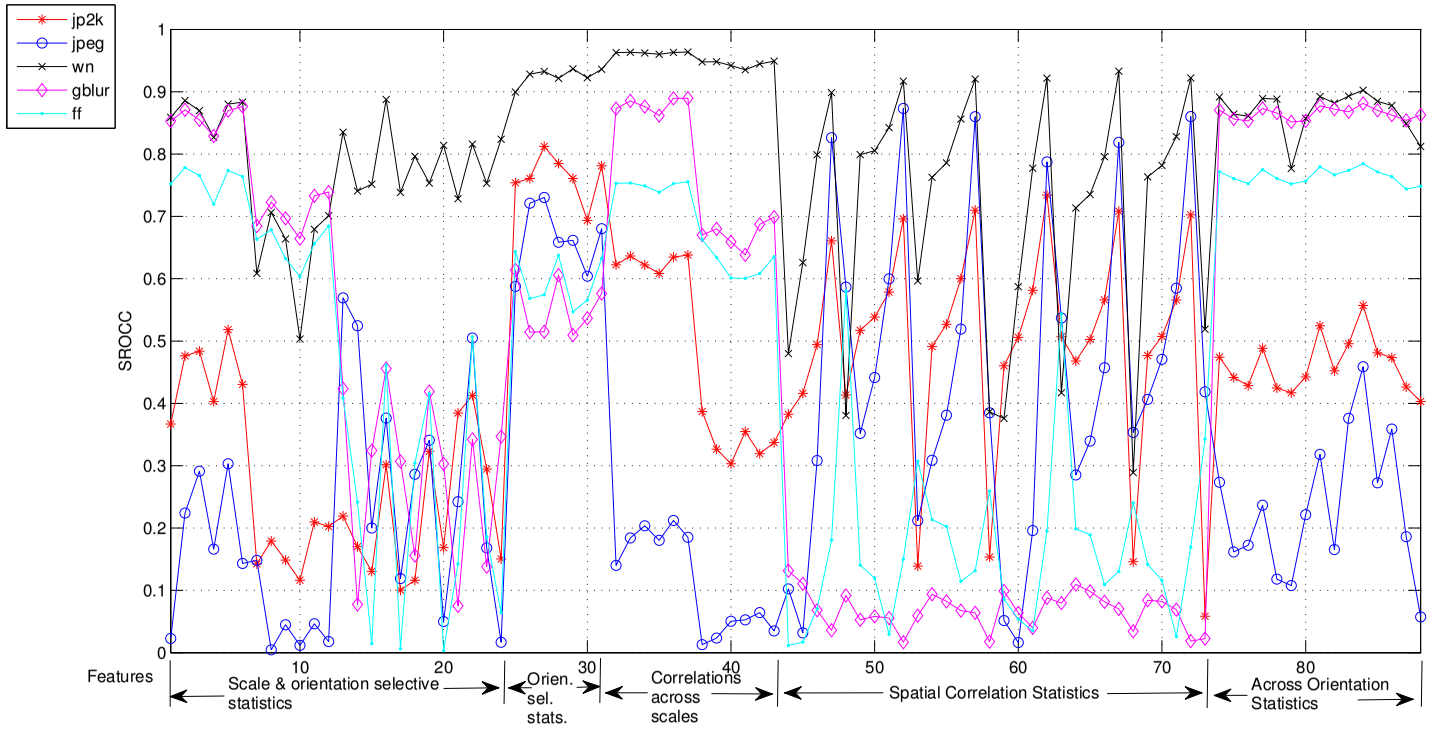


Figure 3.12: Spearman's rank ordered correlation coefficient (SROCC) for each of the features from Table 3.1 on the LIVE image database.

[178]. Both these stages require a calibration process that relates the computed feature to the distortion-class associated with it and the human opinion score associated with it. This calibration is achieved using training, where a set of images whose ground truth class of distortion as well the associated human opinion score (i.e., perceived quality score) is known. Given this training set, we calibrate the two stages of distortion-identification and distortion-specific quality assessment. Once calibrated, DIIVINE is capable of assessing the quality of any distorted image without the need for the reference. Note that the calibration stage also does not require the reference image.

Given a training set of images with known distortion class, spanning the range of distortions (n) the algorithm is being calibrated for, we train a classifier with the true class and the feature vector as inputs. The classifier 'learns' the mapping from feature space to class label, and once calibration is achieved the trained classifier produces an estimate of the class of distortion given an input image (i.e., the feature vector associated with the input image).

Similarly, given a set of training images with known quality scores for each of the n distortion classes, we train n regression modules that map the feature vector to the associated quality score. Since each module is trained specifically for each distortion, these regression modules, once trained, function as distortion-specific assessors of quality, i.e., each trained module will produce an estimate of quality (when given as input an image/feature vector) under the assumption that the image is distorted with that particular distortion. The input image whose quality is to be assessed is passed through each of these

trained distortion-specific quality assessment modules and hence we receive $\vec{q}$, an n-dimensional vector corresponding to the quality estimates from each of these n regression modules.

In our approach, the classifier does not produce a hard classification. Instead, probability estimates are extracted from the classifier, which indicate the confidence that the trained classifier demonstrates in placing the input in each of the n classes. Thus, given an input image/feature vector, the trained classifier produces an n-dimensional vector $\vec{p}$, which represent probabilities of the input belonging to each of the n classes.

Given the two vectors $\vec{p}$ and $\vec{q}$, DIIVINE = $\vec{p}^T\vec{q}$ - i.e., each distortion-specific quality score is weighted by the probability of that distortion being present in the image.

Obviously, one can choose to utilize any classifier and any regression tool to map the feature vectors onto classes/quality scores. In this implementation we utilize a support vector machine (SVM) for classification and support vector regression (SVR) for regression [243, 295]. The choice of SVM and SVR were motivated by the fact that these tools have been shown to perform well on high-dimensional hard classification/regression problems [30]. The interested reader is directed to [30, 243, 295] for detailed explanations of SVMs and SVRs.

We utilize the libSVM package [42] in order to implement the SVM and the SVRs. The kernel used for both classification and regression is the ra-

dial basis function (RBF) kernel, whose parameters are estimated using cross-validation on the training set.

3.3 Performance evaluation

3.3.1 LIVE IQA database

We tested the DIIVINE index on the popular LIVE IQA database [264], which consists of 29 reference images and 779 distorted images that span various distortion categories - JPEG and JPEG2000 compression, white noise, Gaussian blur and a Rayleigh fading channel (fast fading); along with the associated human differential mean opinion scores (DMOS), which are representative of the perceived quality of the image.

Since DIIVINE requires a training stage in order to calibrate the relationship between the extracted statistical features and the distortion category, as well as DMOS, we split the LIVE dataset into 2 non-overlapping sets - a training set and a testing set. The training set consists of 80% of the reference images and their associated distorted versions while the testing set consists of the remaining 20% of the reference images and their associated distorted versions. The classification and regression modules are trained on the training set and the results are then tested on the testing set. In order to ensure that the proposed approach is robust across content and is not governed by the specific train-test split utilized, we repeat this random 80% train - 20% test split 1000 times on the LIVE dataset and evaluate the performance on each of these test sets. The figures reported here are the median of the indices used

for performance across these 1000 train-test iterations[2].

The indices used to measure performance of the algorithm are the Spearman's Rank Ordered Correlation Coefficient (SROCC), the linear (Pearson's) correlation coefficient (LCC) and the Root Mean Squared Error (RMSE) between the predicted score and the DMOS. LCC and RMSE are computed after passing the algorithmic scores through a logistic non-linearity as in [264]. A value close to 1 for SROCC and LCC and a value close to 0 for RMSE indicates superior correlation with human perception. The median SROCC, LCC and RMSE values across these 1000 train-test trials are tabulated in Tables 3.2-3.4, for each distortion category, as well as across distortion categories.

We also report the performance of two FR IQA algorithms - peak-signal-to-noise-ratio (PSNR), and the structural similarity index (SSIM). The former has been used (despite much criticism [94, 310]) as a measure of quality for many years, and the latter is now gaining popularity as a good-yet-efficient assessor of perceived image quality. We also tabulate the performances of several NR IQA algorithms, including original algorithms used to demonstrate the concept of the two-stage framework - the Blind Image Quality Index (BIQI) - BIQI-PURE and BIQI-4D[3], and the two holistic NR IQA algorithms that we have previously discussed - Anisotropy based NR IQA [88] and the BLind Image Integrity Notator using DCT Statistics (BLIINDS) index [235]. The

[2]We use the realigned DMOS scores as recommended in [264] and report results only on the distorted images, as in [264].

[3]The reader is referred to [178] for details on these realizations of the BIQI-framework.

	JP2K	JPEG	WN	Gblur	FF	All
PSNR	0.868	0.885	0.943	0.761	0.875	0.866
SSIM (SS)	0.938	0.947	0.964	0.907	0.940	0.913
BIQI-PURE	*0.736*	*0.591*	*0.958*	*0.778*	*0.700*	*0.726*
BIQI-4D	*0.802*	*0.874*	*0.958*	*0.821*	*0.730*	*0.824*
Anisotropic IQA	*0.173*	*0.086*	*0.686*	*0.595*	*0.541*	*0.323*
BLIINDS	*0.805*	*0.552*	*0.890*	*0.834*	*0.678*	*0.663*
DIIVINE	*0.913*	*0.910*	*0.984*	*0.921*	*0.863*	*0.916*

Table 3.2: Median Spearman's rank ordered correlation coefficient (SROCC) across 1000 train-test trials on the LIVE image quality assessment database. *Italicized* algorithms are NR IQA algorithms, others are FR IQA algorithms.

BIQI realizations are available online [174], and the implementation of the anisotropy measure[4] was obtained from [89]. We implemented the BLIINDS index as described in [235].

It should be clear that DIIVINE performs well in terms of correlation with human perception. Further, DIIVINE improves upon the BIQI realizations, and is superior to the two other holistic NR IQA approaches. Remarkably, DIIVINE also trumps the *full-reference* PSNR, for each distortion separately as well as across distortion categories. However, the most salient observation from the Tables 3.2-3.4 is that the proposed *no-reference* approach is competitive with the *full-reference* SSIM index! This is no mean achievement, since the SSIM index is currently one of the most popular FR IQA algorithms.

Although distortion-identification/classification is not explicitly per-

[4]We note that in [88], the authors mention a correction for JPEG images, which we do not implement here. The variance parameter as suggested is used for NR IQA.

	JP2K	JPEG	WN	Gblur	FF	All
PSNR	0.879	0.903	0.917	0.782	0.880	0.862
SSIM (SS)	0.940	0.947	0.983	0.902	0.952	0.906
BIQI-PURE	*0.750*	*0.630*	*0.968*	*0.800*	*0.722*	*0.740*
BIQI-4D	*0.819*	*0.879*	*0.968*	*0.843*	*0.771*	*0.833*
Anisotropic IQA	*0.130*	*0.083*	*0.490*	*0.469*	*0.420*	*0.187*
BLIINDS	*0.807*	*0.597*	*0.914*	*0.870*	*0.743*	*0.680*
DIIVINE	*0.922*	*0.921*	*0.988*	*0.923*	*0.888*	*0.917*

Table 3.3: Median linear correlation (LCC) across 1000 train-test trials on the LIVE image quality assessment database. *Italicized* algorithms are NR IQA algorithms, others are FR IQA algorithms.

	JP2K	JPEG	WN	Gblur	FF	All
PSNR	11.87	13.60	11.14	11.25	13.33	13.89
SSIM (SS)	8.59	10.11	5.17	7.96	8.74	11.56
BIQI-PURE	*16.54*	*24.58*	*6.93*	*11.10*	*19.48*	*18.36*
BIQI-4D	*14.34*	*15.06*	*6.94*	*9.90*	*17.90*	*15.05*
Anisotropic IQA	*24.65*	*31.40*	*24.41*	*16.19*	*25.44*	*26.68*
BLIINDS	*14.78*	*25.32*	*11.27*	*9.08*	*18.62*	*20.01*
DIIVINE	*9.66*	*12.25*	*4.31*	*7.07*	*12.93*	*10.90*

Table 3.4: Median root-mean-squared error (RMSE) across 1000 train-test trials on the LIVE image quality assessment database. *Italicized* algorithms are NR IQA algorithms, others are FR IQA algorithms.

	JP2K	JPEG	WN	Gblur	FF	All
Class. Acc.(%)	80.00	81.10	100	90.00	73.33	83.75

Table 3.5: Median classification accuracy of classifier across 1000 train-test trials on the LIVE image database.

formed in the two-stage framework used here (recall that we use a probabilistic classification in which the probability of an image belonging to a particular distortion category is estimated), in order to demonstrate that the features are capable of identifying the distortion afflicting the image with high accuracy, in Table 3.5, we list the median classification accuracy of the classifier for each distortion category and the overall accuracy as well. The caveat here is that the actual accuracy of the classifier is not of great import for the proposed approach, since hard classification is never performed. The classification accuracies are reported for completeness.

3.3.2 Statistical Significance Testing

We tabulated the median correlation values of DIIVINE as well as other NR and FR IQA algorithms in the previous section. Although the presented results show some differences in terms of the median correlation, in this section we evaluate if this difference in correlation is statistically significant. Our analysis here is based on the SROCC values across all distortions.

Recall that we computed correlations for each of the algorithms over 1000 test sets. Thus, apart from the median score tabulated before, we have at our disposal the mean SROCC value and the standard error associated with

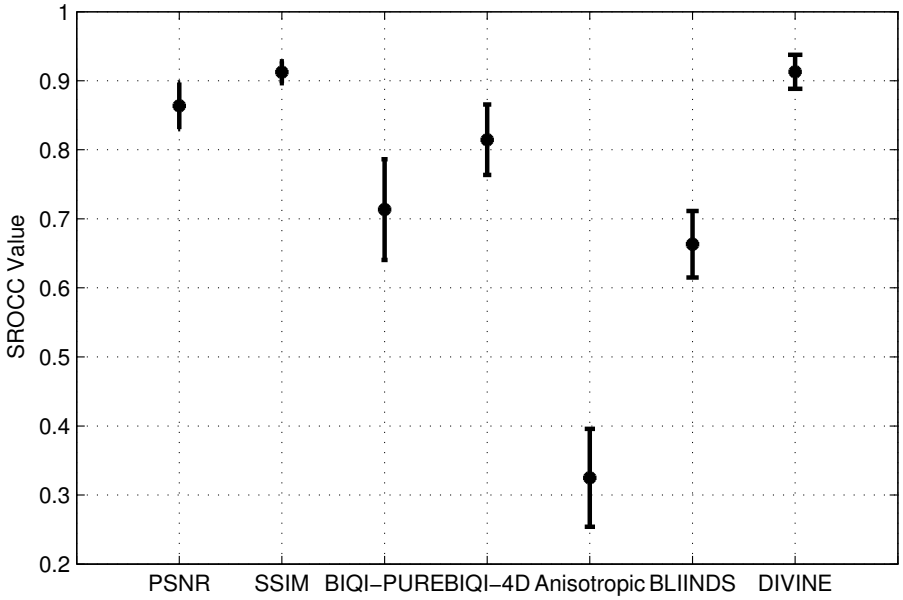


Figure 3.13: Mean SROCC and error bars one standard deviation wide for the algorithms evaluated in Table 3.2, across 1000 train-test trials on the LIVE IQA database.

the 1000 correlation values. In Fig. 3.13, we plot this mean correlation value across the dataset along with error bars one standard deviation wide for each of the algorithms evaluated in Table 3.2.

In order to evaluate statistical significance, we utilize the one-sided t-test between the correlation scores generated by the algorithms across the 1000 train-test trials [266]. In Table 3.6 we tabulate the results of such statistical analysis. The null hypothesis is that the mean correlation of the row is equal to the mean correlation of the column at the 95% confidence level. The alternative hypothesis is that the mean correlation of the row is greater (or lesser) than the

	PSNR	SSIM	*BIQI-PURE*	*BIQI-4D*	*Anisotropic IQA*	*BLIINDS*	*DIIVINE*
PSNR	0	-1	1	1	1	1	-1
SSIM	1	0	1	1	1	1	0
BIQI-PURE	*-1*	*-1*	*0*	*-1*	*1*	*1*	*-1*
BIQI-4D	*-1*	*-1*	*1*	*0*	*1*	*1*	*-1*
Anisotropic IQA	*-1*	*-1*	*-1*	*-1*	*0*	*-1*	*-1*
BLIINDS	*-1*	*-1*	*-1*	*-1*	*1*	*0*	*-1*
DIIVINE	*1*	*0*	*1*	*1*	*1*	*1*	*0*

Table 3.6: Results of the one-sided t-test performed between SROCC values. A value of '1' indicates that the algorithm (row) is statistically superior to the algorithm (column). A value of '0' indicates statistical equivalence between the row and column, while a value of '-1' indicates that the algorithm (row) is statistically inferior to the algorithm (column). *Italicized* algorithms are NR IQA algorithms, others are FR IQA algorithms.

mean correlation of the column. Table 3.6 indicates which row is statistically superior ('1'), statistically equivalent ('0') or statistically inferior ('-1') to which column.

From Table 3.6, it is obvious that DIIVINE is statistically better than other no-reference approaches to IQA. Further, *DIIVINE is statistically superior to the full-reference PSNR.* This is a significant result indeed, for we are unaware of any NR IQA algorithm that is not only capable of assessing quality across many distortion categories, but also performs statistically better than the full-reference PSNR. Indeed, DIIVINE, which predicts perceived quality given ONLY the distorted image produces correlations with human subjective judgments at a level that is *statistically indistinguishable from the full-reference structural similarity index (SSIM)* that needs both the reference and distorted image in order to assess quality! This suggests that one can safely replace the FR SSIM with the NR DIIVINE without any loss in performance, provided

that the distortions encountered are well-represented by the dataset used to train DIIVINE (here - the LIVE IQA database).

3.3.3 Database Independence

Since NR IQA algorithms are generally trained and tested on various splits of a single dataset (as described above), it is natural to wonder if the trained set of parameters are database-specific. In order to demonstrate that the training process is simply a calibration, and once such training is performed, DIIVINE is capable of assessing the quality of any distorted image (from the set of distortions it is trained for) we evaluate the performance of DIIVINE on an alternate database - the TID2008 [221].

The TID database consists of 25 reference images and 1700 distorted images over 17 distortion categories. Of these 25 reference images only 24 are natural images and we test our algorithm only on these 24 images. Further, of the 17 distortion categories we test DIIVINE only on those categories it has been trained for - JPEG, JPEG2000 compression (JP2k), Additive white noise (WN) and Gaussian Blur (blur)[5]. In order to evaluate DIIVINE on the TID database, we train the parameters of DIIVINE using the entire LIVE IQA database as described previously. The trained model is then tested for its performance on the TID database. In Table 3.7, we tabulate the SROCC values obtained for such testing for each distortion as well as across distortion

[5]Although DIIVINE has been trained for FF, the JP2k transmission loss distortion on the TID database does not correspond to this kind of fading channel model and hence is not considered here.

	JP2K	JPEG	WN	Gblur	All
PSNR	0.825	0.876	0.918	0.934	0.870
SSIM (SS)	0.963	0.935	0.817	0.960	0.902
DIIVINE	*0.924*	*0.866*	*0.851*	*0.862*	*0.889*

Table 3.7: Spearman's rank ordered correlation coefficient (SROCC) on the TID2008 database. *Italicized* algorithms are NR IQA algorithms, others are FR IQA algorithms.

categories. Further, we also list the performance of the FR PSNR and SSIM for comparison purposes. It is clear from Table 3.7 that the performance of DIIVINE is NOT database dependent and that once trained DIIVINE is capable of assessing the quality of images across the distortions that it is trained for.

3.3.4 Computational Analysis

Although DIIVINE was not developed under the constraint of real-time analysis of images, given that the performance of DIIVINE is as good as leading FR QA algorithms, its computational complexity is relevant when one considers applications of DIIVINE. Hence, it is prudent to perform an informal analysis of the computations needed to predict the quality of an image without a reference using DIIVINE.

An unoptimized MATLAB code takes approximately 60 seconds to produce a quality estimate on a 1.8 GHz processor with 2 GB of RAM running Windows XP and MATLAB R2008a for a 512×768 image. The amount of time taken for training the SVM/SVRs is negligible as is the time taken to

Step	Percentage of Time
Steerable Pyramid Decomposition	2.52
Divisive Normalization	10.83
Or. & Scale selective Statistics	0.10
Orientation selective Statistics	0.48
Across scale Correlations	8.42
Spatial Correlation	69.72
Across Orientation Statistics	7.92

Table 3.8: Informal complexity analysis of DIIVINE. Tabulated values reflect the percentage of time devoted to each of the steps in DIIVINE.

predict the quality by the trained classifier/regressors compared to that of feature extraction. In Table 3.8 we tabulate the percentage of time devoted to each of the steps in DIIVINE.

As is clear from Table 3.8, spatial correlation statistics occupy a considerable chunk of the processing time. This is primary because constructing the 2D PDFs needed for various spatial shifts is a computationally intensive process. One would imagine that implementing this efficiently in compile-able code (such as C) would cut down the time needed considerably. Further, the steerable pyramid decomposition in this version of DIIVINE is performed using the MATLAB toolbox from the authors [268], without using MEX code as recommended. Given that there exists C code for the same, it is not wrong to suppose that the time take for this section may also be reduced drastically. Similar arguments hold for the divisive normalization process. The across-orientation statistics and the across scale correlations are based on windowed structural correlation computation, whose current implementation is in MATLAB. Recently, however, faster real-time implementations of such windowed

correlations have been made available, which would reduce the computation associated with these steps as well [54].

Thus, it seems that DIIVINE can be re-coded efficiently in order to achieve close-to-real-time (if not real-time) performance. Thus, application of DIIVINE should not suffer owing to its complexity.

3.4 Discussion and Conclusion

We proposed a no-reference (NR)/blind image quality assessment (IQA) framework and integrated algorithm based on natural scene statistics, that assesses the quality of an image without need for a reference across a variety of distortion categories. This algorithm - the Distortion identification-based Image Verity and INtegrity Evaluation (DIIVINE) index - utilizes the previously proposed two-stage framework which first identifies the distortion present in the image and then performs distortion-specific quality assessment to provide an ostensibly distortion-independent measure of perceptual quality, using extracted natural scene statistic features. We detailed the statistical features extracted, along with motivations drawn from vision science and image processing, and demonstrated that the DIIVINE index correlates well with human perception of quality. We undertook a thorough analysis of the proposed index on the publicly available LIVE IQA Database, and showed that the proposed measure is statistically superior to other NR IQA algorithms that function across distortion categories. Further, we compared the performance of DIIVINE with two standard full-reference QA algorithms: the peak signal-

to-noise-ratio (PSNR) and the single-scale structural similarity index (SSIM). We showed that DIIVINE is *statistically superior* to the FR PSNR and *statistically indistinguishable* from the FR SSIM. To the best of our knowledge, DIIVINE is the only IQA algorithm that not only assesses quality across a range of distortions, but also correlates with human perception judgments at a level that is statistically equivalent to good FR measures of quality. Finally, we demonstrated that DIIVINE performance is database-independent and can easily be extended to distortions beyond those considered here, and performed an informal complexity analysis.

The proposed approach is modular, and can easily be extended beyond the set of distortions considered here. Importantly, DIIVINE does not compute specific distortion features (such as blocking), but instead extracts statistical features which lend themselves to a broad range of distortion measurements. Future work will involve increasing the subset of distortions beyond those considered here, in an effort to further relax any distortion dependence. A software release of DIIVINE has been made available online: http://live.ece.utexas.edu/research/quality/DIIVINE_release.zip.

Chapter 4

Perceptually Optimized Blind Repair of Natural Images

Image repair refers to the process of correcting one or more possibly different types of distortions afflicting an image. The general purpose image repair problem is formulated as:

$$\mathbf{y} = \mathbf{H} \cdot f(\mathbf{x}) + \mathbf{n} \tag{4.1}$$

where $\mathbf{y}$ is the observed distorted image, $\mathbf{x}$ is the original pristine image that we seek to recover, $\mathbf{n}$ is the additive noise, $f(\cdot)$ is a local non-linearity and $\mathbf{H}$ is a matrix that models multiplicative distortion (e.g., a low-pass filter) [11, 15]. The model in (4.1) is not restricted to the spatial domain (where the vectors $\mathbf{x}$ and $\mathbf{y}$ would be columnized versions of the 2D image) and we do not assume that the models for image repair are limited to the spatial domain.

The general image repair problem is ill-posed, and in order to solve the problem, certain assumptions are usually made about the structure of $f, \mathbf{H}$ and $\mathbf{n}$. For example, for image denoising, fix f to be an identity transform, and $\mathbf{H} = \mathcal{I}$, where $\mathcal{I}$ is the identity matrix, and assume a distribution and correlation structure on the noise $\mathbf{n}$ [222]. For deblurring (deconvolution),

assume a zero-mean noise model with known variance, then estimate $\mathbf{x}$ from the observed $\mathbf{y}$ [140] and so on.

Image repair algorithms have been broadly partitioned into blind and non-blind classes. Blind algorithms do not assume prior knowledge of the distortion parameters, while non-blind models assume that the parameters are known. Given the ill-posed nature of the problem, there has been more activity and success on various non-blind image repair problems than on blind image repair problems [29, 49]. While the general field of image repair has seen quite a bit of research, especially on single distortion problems such as denoising [29, 49], deconvolution [119, 140] and deblocking [27, 277], the general purpose blind image repair problem, where the specific distortion(s) afflicting the image are unknown, has been little studied.There has been some work on dealing with two image distortions simultaneously (and blindly, if the blur/noise parameters are unknown), the classic example being the image restoration problem [119] of simultaneously deblurring and denoising an image. The common theme of these approaches is that it is known *a priori* what the distortions are that afflict the image. A blind algorithm then seeks to discover the parameters of the distortion (noise, blur) and then ameliorate them.

We take a different approach to the problem of general purpose image repair. We begin by assuming that the distortion(s) (if any) afflicting a given image are unknown and possibly multiple, although they are assumed to come from a finite population of possible image distortions. We refer to such a problem framework as *distortion blind*. We also recognize that for specific

image distortions of general interest, there exist algorithms that ostensibly correct that particular distortion reasonably well. This is not always the case, of course; for example, image restoration (deblur and denoise simultaneously) is a particularly difficult inverse problem that requires precise modeling to achieve worthwhile results. This is often impossible given that blur nearly always arises from a non-linear process. In any case, in this work we do not attempt to improve upon the state of the art of any type of image repair problem. Rather, we propose the new idea of preprocessing the image to determine the distortion(s) afflicting it and any unknown parameters of the distortion(s), then once done, ameliorate these using the best algorithm available. Of course, if the distortion identification stage is particular effective (e.g., by better parameter estimation), then performance may be notably improved.

Thus we define the task of a general purpose image repair algorithm to be agglomeration, i.e., automatically deploy any of multiple high-performing image repair algorithms towards achieving seamless general purpose image repair across a wide variety of distortion types. Such a general purpose image repair algorithm should perform as well as the best algorithms on each included subclass of distortion (since it embodies these algorithms in its architecture). As mentioned, given the ill-posed nature of many inverse image repair problems, it is natural that repair algorithms (even the best of them) would fail in certain situations. In these cases we further posit that, the general-purpose image repair algorithm should be able to detect the failure and act on it so that the repaired image is given the best *perceptual* quality at the output, thereby

rejecting failures by the internal repair algorithms. Further, given that each of the subclass repair algorithms may also introduce new artifacts (e.g., deblocking can introduce blur in the image), the general-purpose algorithm should enable iterative distortion correction within the set of subclasses that it encompasses. Finally, if some of the best repair algorithms are non-blind, the general-purpose image repair algorithm could include blind parameter estimation modules so that these non-blind algorithms operate using these estimated parameters, towards solving a class-specific blind image repair.

Here we propose both a general design framework – the GEneral-purpose No-reference Image Improver (GENII) – for general purpose distortion-blind image repair as well as an example working model and algorithm dubbed GENII-1. Our framework and exemplar models are based on using natural scene statistics (NSS) [222, 231, 270] to identify distortions by type and severity. The specific exemplar model, GENII-1, is capable of restoring images distorted by additive noise, Gaussian blur, JPEG compression or JPEG2000 compression, without knowing in advance which (if any) of the distortions impairs the image, or the parameters of the distortion. Given a distorted image, the algorithm uses natural scene statistic (NSS) features to first identify whether the image has been distorted, and if so, identify (a) the distortion that afflicts the image, (b) the associated distortion parameter (e.g., noise variance) and (c) the perceptual quality of the image. The algorithm then proceeds to apply an appropriate off-the-shelf image repair algorithm based on the identified distortion category. The perceptual quality of the repaired image so

obtained is repeatedly evaluated and the general-purpose image repair loop continues until a maximum level of objectively determined perceptual quality is obtained. Thus our model seeks to guarantee that the final repaired image will not only be distortion-reduced, but will also present the best possible perceptual quality. The entire process is completely blind both to the distortion type and the distortion parameters. The only information available to the algorithm is the fact that the image it is trying to repair belongs to the category of natural images[1] and that the distortion (if any) belongs to one of the multiple diverse distortions.

In order to achieve these goals, we use a realization of our previously proposed two-stage framework for image quality assessment [178] that first identifies the distortion that afflicts the image [177] and then proceeds to assess quality (Chapter 3). We have developed two realizations of this two-step framework – one in the wavelet domain [179] (Chapter 3) and the other in the spatial domain [169] (not described in this dissertation) – both capable of accurate distortion-identification as well as blind parameter estimation and blind image quality assessment. Either can be combined with off-the-shelf distortion-specific image repair algorithms to perform distortion-blind general-purpose, image repair. We are unaware of any approach that takes this approach to the general image repair problem and to the best of our knowledge, the proposed model is the first of its kind.

[1]Natural images are those images captured by a camera, and do not include computer-generated renders of the visual world.

Before we proceed,we define the terminology used in the rest of this article. A distortion *class* refers to a particular kind of distortion afflicting the image, e.g., blur or noise. A distortion *type* refers to a particular type of a distortion class, eg., Gaussian noise, or spatially invariant Gaussian blur. Image repair refers to eliminating distortions arising from any one of the multiple classes of distortions. An exemplar implementation of the GENII framework (GENII-1) ameliorates distortions arising from any one of multiple distortion types.

4.1 Distortion Blind Image Repair

Our approach to general purpose image repair is summarized as follows. Given an input (possibly) distorted image, we first extract statistical DIIVINE features from the image, then use these features to attempt to identify the distortion afflicting the image. Once a distortion class has been posited, the same features are used to predict the perceptual quality of the image using the second stage of quality assessment. If the predicted quality of the image lies above a certain threshold, then the quality of the image is deemed to be high enough that repair does not need to be performed; in which case, the algorithm halts, yielding as output the input image. If the predicted quality falls below this threshold, the algorithm continues.

Given identified distortion(s), the same DIIVINE features are used to perform blind parameter estimation for the corresponding image repair problem, for example, these features might predict the noise variance in the image

if the distortion is predicted to be Gaussian noise. The algorithm then proceeds to invoke the appropriate image repair algorithm, providing as input to this algorithm the distorted image to be repaired and the associated parameter(s) that the off-the-shelf (possibly non-blind) algorithm may require. The repaired intermediate image so obtained is then passed back into the loop in order to evaluate quality and identify distortion. This loop continues until the obtained intermediate image has the highest possible quality or if a finite number of repair iterations have been performed. While this procedure does not guarantee convergence, we have not found any example among the 4000 distorted images that we tested on (see below) that produced convergence issues. Of course, the introduction of a suitable stopping criterion belays this question. Furthermore, the algorithm determines whether the repaired image has a higher predicted visual quality than the input distorted image, and only outputs the repaired image if it has a higher visual quality, thereby avoiding unpalatable distortions that the repair algorithm may introduce as well as compensating for the failure of the repair algorithm. An illustration of this general purpose image repair scheme is diagrammed in Fig. 4.1.

Note that the approach that we have proposed is highly modular in the sense that any repair scheme can be replaced by another repair scheme deemed to be more effective than the one on the system. Further, the model that we have described may deploy either blind or non-blind repair algorithms. When non-blind, our features can be used to predict repair algorithm parameters.

In this chapter, we demonstrate the efficacy of our general-purpose

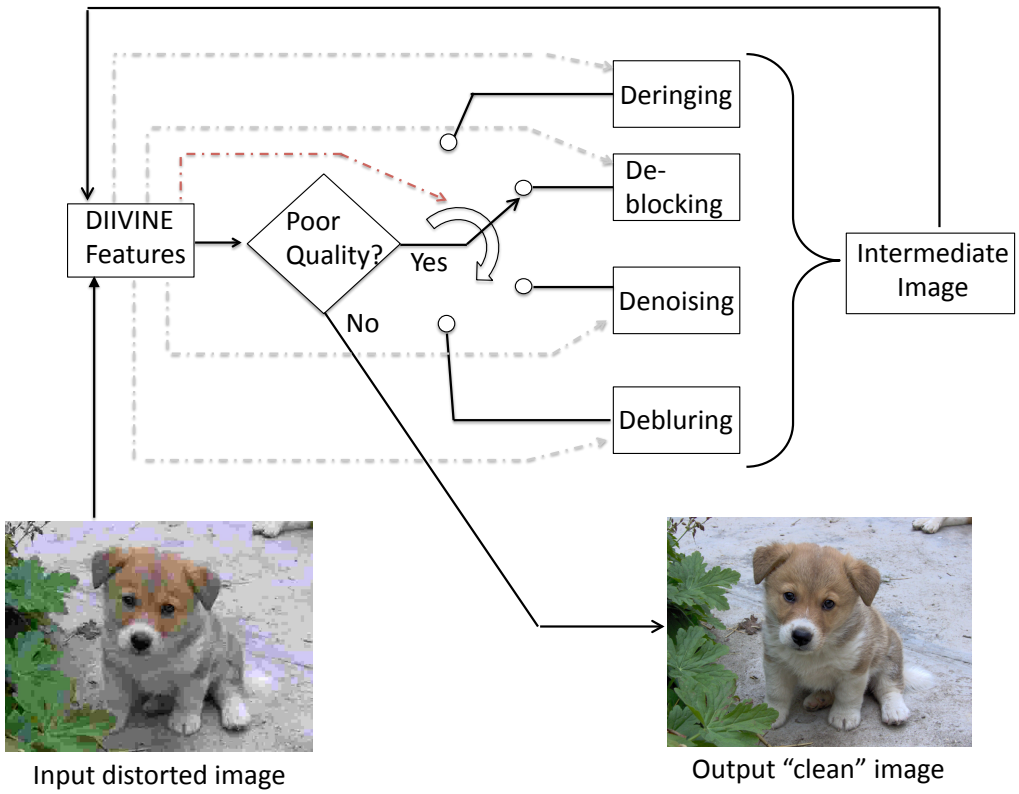


Figure 4.1: An illustration of the GENII framework. DIVIINE features are used to predict the distortion class, the visual quality, and the distortion parameters that may serve as inputs to a possibly non-blind repair algorithm. The intermediate repaired image is fed back to the system until the best possible quality is achieved at the output.

image repair framework using four distortion repair subtypes. However, the GENII framework encompasses the possibility of more extensive implementations capable of handling more than the four distortion classes considered here, diverse subtypes of the distortions (different types of blur, noise, missing data, or other artifacts), and appropriate combinations of distortion. The resulting algorithm, GENII-1, is responsive to these distortion although blind to which (if any) of these distortions occur. It is limited since it is restricted to these four distortions and is not trained to identify combinations of them or equipped to repair multiple coincident distortions. Currently this capability is limited by existing distortion databases and by availability of human judgements of multiply distorted images.We have planned future image databases to address these possibilities.

4.1.1 Image Repair Algorithms

As described in the introduction, GENII uses a two-stage framework, where once the distortion is identified, an off-the-shelf image repair algorithm is deployed to conduct distortion-specific image repair. Now we shall summarize the exemplar repair algorithms that we have decided to use to demonstrate the principle of general purpose image repair, in our prototype implementation GENII-1. The algorithms were chosen since they were either readily available online or were easy to implement, have a previously demonstrated high-level of performance, and adequate computational efficiency without sacrificing performance. The modularity of our model implies that any one of these algorithms

could be substituted for by a suitable alternative.

4.1.1.1 Deblocking

We use the simple algorithm proposed in [196], which iteratively applies JPEG compression at the quality level at which the distorted image was compressed, to shifted versions of the distorted image. The resulting collection of images is averaged to produce a final deblocked image. The premise behind this approach is explained in [196], and we have found that the algorithm efficiently reduces blocking artifacts in a perceptually satisfying manner. The input parameter required, since the algorithm is not blind, is the quality factor at which the image was compressed. This can be read from the JPEG header, or it can be estimated, as we demonstrate below.

4.1.1.2 Deringing

We use the trilateral filter described in [306] to remove ringing artifacts from the image. The trilateral filter is an extension of the bilateral filter [75], which first computes a texture map from the gradient information and then filters the image using a locally adaptive filtering procedure, where the filter kernels are functions of the image intensity and the textural information at each location. We tried other deringing approaches (for example, the one in [197]). However, while these approaches reduced ringing artifacts, the images produced had poorer quality than the distorted image, both by visual inspection and by quantitative QA analysis [319], while the trilateral filter produced

higher quality images. This algorithm does not need any input parameter, i.e., the algorithm is blind.

4.1.1.3 Denoising

We use the Block matching 3D (BM3D) algorithm for denoising [61]. The BM3D algorithm operates as follows. Given a distorted image, with known noise variance, a set of groups of 2D image patches are created via block-matching to produce a 3D group of image patches, each of which are then denoised in a sparse transform domain using a popular wavelet shrinkage based approach [71]. These denoised patches yield a basic estimate of the denoised image, which is then used to perform re-grouping, followed by collaborative Wiener filtering, where the 'collaboration' is between the image patches in the group. The algorithm, which was designed for Gaussian noise, is not blind and the input parameter is the noise variance.

4.1.1.4 Deblurring

The approach proposed in [140] is used for deconvolution. Local image gradients are modeled using a heavy-tailed distribution, which forms a natural image prior. A *maximum a posterior* (MAP) problem is solved using the iterative re-weighted least squares (IRLS) approach [166]. The algorithm requires the blur kernel as prior information to be able to perform deblurring. Since we consider spatially invariant Gaussian blur, the parameter to be estimated is simply the variance of the blur kernel.

4.2 Implementation and Performance Evaluation

Since there are a variety of stages involved in the exemplar image repair algorithm, GENII-1, we evaluate each stage individually, then the overall performance. Further, in order to demonstrate the robustness the GENII concept, instead of evaluating it only on a standard set of images, we perform a more complete analysis on the repair performance of GENII-1 on a much larger database. To this end, we created a large database of distorted images spanning a wide range of distortion levels encompassing the four types that GENII-1 has been designed to repair : JPEG2000 compression (JP2K), JPEG compression (JPEG), additive white noise (WN) and Gaussian blur (Blur).

A total of 300 reference images from the Berkeley image segmentation database [161] were distorted at 10 different degrees of severity for each distortion type to produce a total of 12000 distorted images (3000 per class). JPEG compression was implemented using MATLAB's `imwrite` command; JPEG2K was implemented using the Kakadu encoder [282]; zero-mean WN was added to the image using MATLAB's `imnoise` command; and Blur was simulated using a Gaussian kernel to filter the image. The various control parameters for these distortions and the ranges of these parameters for the simulated distortion levels are listed in Table 4.2. The distortion levels were uniformly sampled on a log-scale between the minimum and maximum parameter values. Fig. 4.2 plots some sample reference and associated distorted images to give a sense of the distortion levels created.

In the discussion above, we described how GENII-1 performs distor-

(a) Reference image

(b) Reference crop

(c) Distorted crop: JP2K

(d) Distorted crop: JPEG

(c) Distorted crop: WN

(d) Distorted crop: Blur

Figure 4.2: Sample simulated distorted images (crops) from the Berkley image segmentation database [161] .

Distortion type & Parameter	Min. Value	Max. Value
JP2K (bit-rate)	0.05	0.25
JPEG (quality parameter)	7.5	20
WN (σ^2 of filter)	0.001	0.05
Gblur (σ of filter)	1	10

Table 4.1: Distortion parameters and their minimum and maximum values used for inducing distortions.

tion identification and blind parameter estimation. To achieve this, GENII-1 requires a training phase in which the extracted features are mapped onto the associated distortion type as well as to the distortion parameters. In order to train our algorithm, we split the above database randomly, based on image content, such that 200 reference images (and the associated 8000 distorted images) are used for training and the remaining 100 reference images (and the associated 4000 distorted images) are used for testing. All results to follow are reported on this testing set. This train-test split procedure ensures that there is no content overlap between the training and test sets. Next, we summarize the training procedure.

4.2.1 Training the Model

4.2.1.1 Classification

A multi-class support vector machine (SVM) [42, 295] is trained to classify the distorted images into one of four distortion types using DIIVINE features as inputs and the labels associated with the distortion types as the outputs. The parameters of the SVM are set via cross-validation. Once trained, when fed with DIIVINE features, the SVM returns a distortion type and a

probability distribution over all distortion types which corresponds to the predicted type and the confidence associated with the classification respectively. This predicted distortion type is used to select the right image repair algorithm.

4.2.1.2 Quality Assessment

The confidence associated with the prediction (probability estimates) are used in conjunction with regression modules to accomplish quality assessment. This procedure is described in great detail in [178, 179]. Supposing n distortion types (in GENII-1, $n = 4$), n regression modules (support vector regression (SVRs) [295]) are trained, taking as input DIIVINE features, and then regressed on to (known) quality scores for each of the distortion types independently. Since human opinion scores are not available for the database that we created, we instead use the multi-scale structural similarity (MS-SSIM) index [319] to supply quality scores. MS-SSIM produces quality predictions that correlate quite well with human judgments of quality of images impaired by these and many other types of distortions [264]. As such it is a useful proxy for human opinion scores. However, MS-SSIM correlates non-linearly with human judgments of quality. Therefore, instead of using the MS-SSIM scores directly, the MS-SSIM scores are remapped to human opinion scores obtained from the LIVE IQA database [265]. These remapped scores have a range of [0, 100], where '0' is the best possible subjective quality. This procedure is detailed in Appendix A. In order that the distinction between the MS-SSIM

scores and the remapped MS-SSIM scores is clear, the remapped scores are labelled MS-SSIM$_D$.

During the test phase, given an input distorted image, the algorithm uses the classifier module to produce probability estimates for the types p_i, and for each distortion type produces a quality scores q_i. The final DIIVINE predicted quality score is then $\sum_i p_i q_i$.

4.2.1.3 Parameter Estimation

The repair algorithms used by GENII-1 are mostly non-blind and require as input certain parameters described in Section 4.1. To predict these repair parameters, for each distortion type and for each image in the training set, we use DIIVINE features to train a regression module (SVR [295]) to perceptually optimize the parameter of interest (e.g., variance of the blur kernel). Given the input test image, the image is first classified by the distortion type and the appropriate (trained) regression module is queried to output the estimated parameter of that distortion. This parameter and the appropriate repair algorithm are used to produce a repaired image.

Denoising using BM3D [61] is handled in a slightly different manner to promote an improved perceptual result. Although BM3D requires noise variance as the input, when fed with the actual noise variance in the distorted image, the algorithm tends to over-smooth the image resulting in lower quality than when the algorithm is fed with a different (albeit incorrect) input noise variance [171]. For example, Fig. 4.3 shows a noisy image, its repaired

(a) (b) (c)

Figure 4.3: Accurate noise variance as input to the algorithm in [61] produces poorer quality denoised images: (a) Noisy Image ($\sigma = 0.0158$, MS-SSIM$_D$ = 107.26), (b) Denoised with $\sigma = 0.0158$ (MS-SSIM$_D$ = 64.00) and (c) Denoised with $\sigma = 0.0040$ (MS-SSIM$_D$ = 53.82).

version using the correct noise variance input to BM3D, and the perceptually optimized approach detailed below, which uses a different variance parameter as input to BM3D. BM3D tends to oversmooth images when provided with the actual noise variance, and is capable of producing better quality images when supplied with a different input parameter. We modify BM3D to improve visual quality using an important new aspect of GENII: perceptually optimized training of distortion repair parameters.The training procedure for the denoising module is modified in the following way.

To maximize the visual quality of the denoised image, we use the training procedure outlined in [171]. Specifically, during the training phase, the distorted image is denoised multiple times using BM3D with different input noise variances. The resulting repaired image is then quality-assessed using the perceptually relevant MS-SSIM [319] index. For each distorted image, the value of the input noise variance that maximizes the visual quality (as gauged by MS-SSIM) is the parameter value to train the noise parameter regression

module. In summary, instead of using the actual noise variance as input to the regression module, a different perceptually optimized value is chosen.

Note that this training procedure is specific to the BM3D denoising algorithm and may not be needed for other algorithms that might be used to replace BM3D in an improved GENII implementation. On the other hand, perceptual optimization of an image repair parameters is a powerful option.

In summary, we train one classification module of GENII-1 that outputs a distortion type as well as a probability distribution over types, four regression modules that output quality scores independently for each of the classes, and four parameter estimation regression modules which output the appropriate distortion parameter, all of which utilize as input only natural scene statistic DIIVINE features. During the test phase, the DIIVINE features are extracted just once and from them all of the necessary outputs are produced using the trained modules to (1) classify the image, (2) assess quality, and (3) estimate distortion parameters. The predicted distortion type and the distortion parameters are used to perform image repair. As we noted before, this entire process can be repeated until a maximum quality is achieved.

In the sections that follow, we report results on only a single pass through the system i.e., the output repaired image is not fed back into the loop for further correction. However, we later demonstrate an example of iterative image repair as well. Figure 4.4 shows an example of the operation of the GENII-1, where deconvolution was performed.

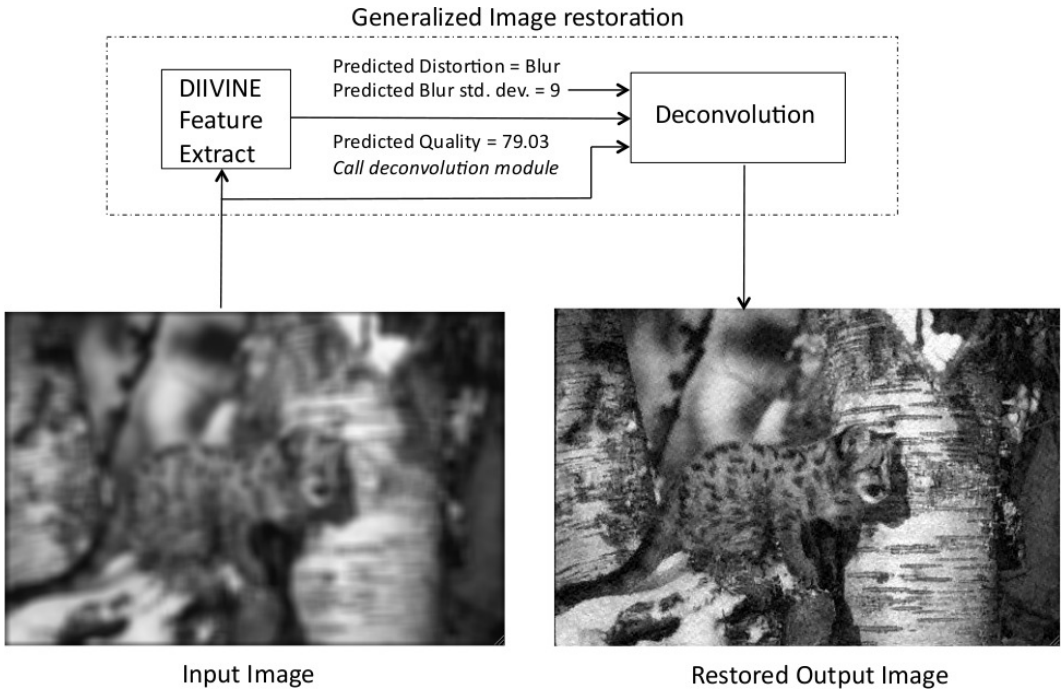


Figure 4.4: Illustration of the operation of GENII-1 using DIIVINE features extracted from the input image. These features are used to identify the distortion, predict the quality and estimate the blur kernel standard deviation. The distorted image and the blur kernel are then fed to the appropriate repair scheme – deconvolution – to produce the output repaired image.

	JP2K	JPEG	WN	Blur	All
DIIVINE	98.10%	98.20%	100%	99.30%	98.90%

Table 4.2: Classification accuracies of DIIVINE.

4.2.2 Performance Evaluation

Next, we evaluate the specific instantiation of our general purpose image repair framework, GENII-1, on the test image database. Specifically, we evaluate the accuracy achieved in predicting distortion type, intermediate image quality, distortion parameters and the final improvement in visual quality obtained after repair.

4.2.2.1 Classification and Quality Assessment

Table 4.2 reports classification accuracy for each distortion type as well as the overall classification accuracy over the 4000 distorted images using both DIIVINE features. Excellent classification accuracy was achieved; there was very little confusion between distortion types ($< 1\%$) and hence, for brevity, we do not report these numbers here.

We computed the Spearman's rank ordered correlation coefficient (SROCC) between the predicted quality scores from DIIVINE and the objective predicted by MS-SSIM$_D$ on the test images across all distortions. We also computed the commonly used *full-reference* peak signal-to-noise ratio (PSNR) as an additional comparison. The SROCC values relative to MS-SSIM$_D$ that were observed were – PSNR = 0.8066, DIIVINE = 0.9308. As expected, DIIVINE correlates much better with the perceptually relevant MS-SSIM$_D$ than

does PSNR.

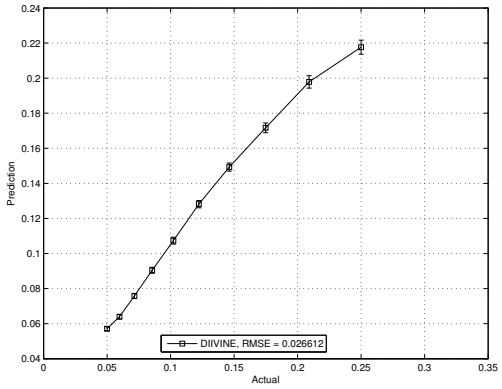
4.2.2.2 Parameter Estimation

Fig. 4.5 plots the mean estimated parameters for each distortion type as a function of the actual input parameters. The associated standard error bars across the 100 different contents in the test set are given for DIIVINE. The figure also lists the root mean-squared-error (RMSE) between the actual value and the predicted value. DIIVINE does a good job of predicting the distortion parameter, and hence their predictions can be used as inputs to non-blind repair algorithms.

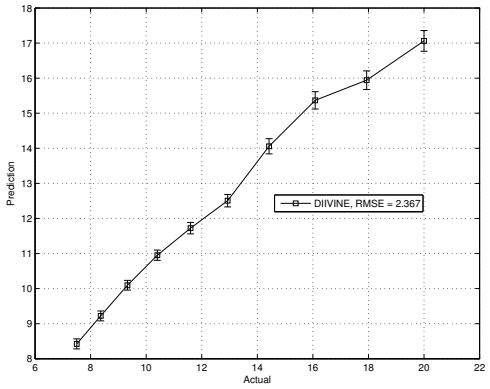
Note that in GENII-1, for the denoising task, we do not actually use the predicted noise variance as the parameter for BM3D, for the reasons explained earlier. Also, the deringing algorithm used does not require any input parameter. The plot simply demonstrates that the DIIVINE framework is capable of predicting these distortion parameters with a high degree of accuracy.

4.2.2.3 Image repair

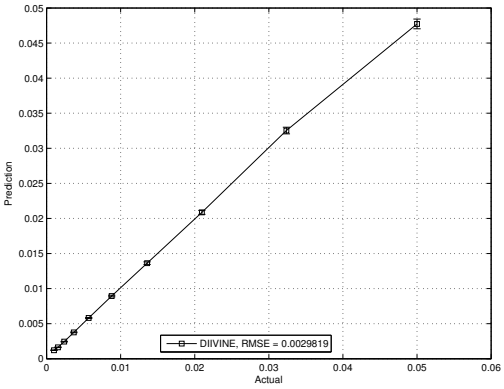
Having demonstrated that DIIVINE features are capable of classifying images according to distortion types, quality assessment, and blind parameter estimation, we now demonstrate how these stages can be combined to fully realize GENII-1 using off-the-shelf algorithms for image repair. Since we are unaware of any other general purpose image repair technique similar to GENII-1, any comparison is impossible. Hence, we report the mean in-



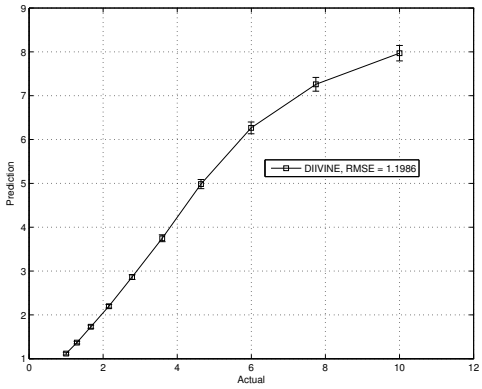
(a) JP2K: bit rate



(b) JPEG: quality parameter



(c) WN: noise standard deviation



(d) Blur: standard deviation of blur kernel

Figure 4.5: Parameter estimation using DIIVINE: Plots of (mean) predicted vs. actual parameters and the standard error bars of distortions considered here. Each subfigure indicates the distortion type and the root mean-squared-error (RMSE) between the actual and predicted values.

crement in MS-SSIM$_D$ quality (and the standard error bars) after the repair process over the baseline MS-SSIM$_D$ quality of the distorted image in Fig 4.6 for each distortion type, and across distortion classes. The results in Fig. 4.6 are for a single-pass of GENII-1, which only repairs the image once based on the distortion category and then calls the appropriate repair algorithm. This single-pass implementation of GENII-1 also checks the image quality at the output and returns the image (repaired or input distorted) having the higher quality, as predicted by the image quality index being used (DIIVINE) . This quality check guarantees the best quality at the output, and accounts for distortions possibly introduced in the repair process which may have reduced the perceptual quality (although the original distortion may have been repaired, e.g., deblocking leading to blur).

For strictly comparison purposes, Fig. 4.6 also plots the quantitative quality improvement obtained when using each image repair algorithm with perfect knowledge of the distortion (i.e., without the classifier stage), and perfect knowledge of the input distortion parameters (i.e., using non-blind algorithms). Although such a baseline is unfair to the algorithm being evaluated, it provides insights into the performance of the proposed approach.

Figure 4.6 indicates that GENII-1 performs quite well predicting and ameliorating the distortions present in the image. Note that the improvement in quality that GENII-1 delivers is limited by the performance of the repair modules that it uses. A better repair algorithm will lead to greater increases in visual quality, as evidenced by the large gains in MS-SSIM$_D$ obtained for

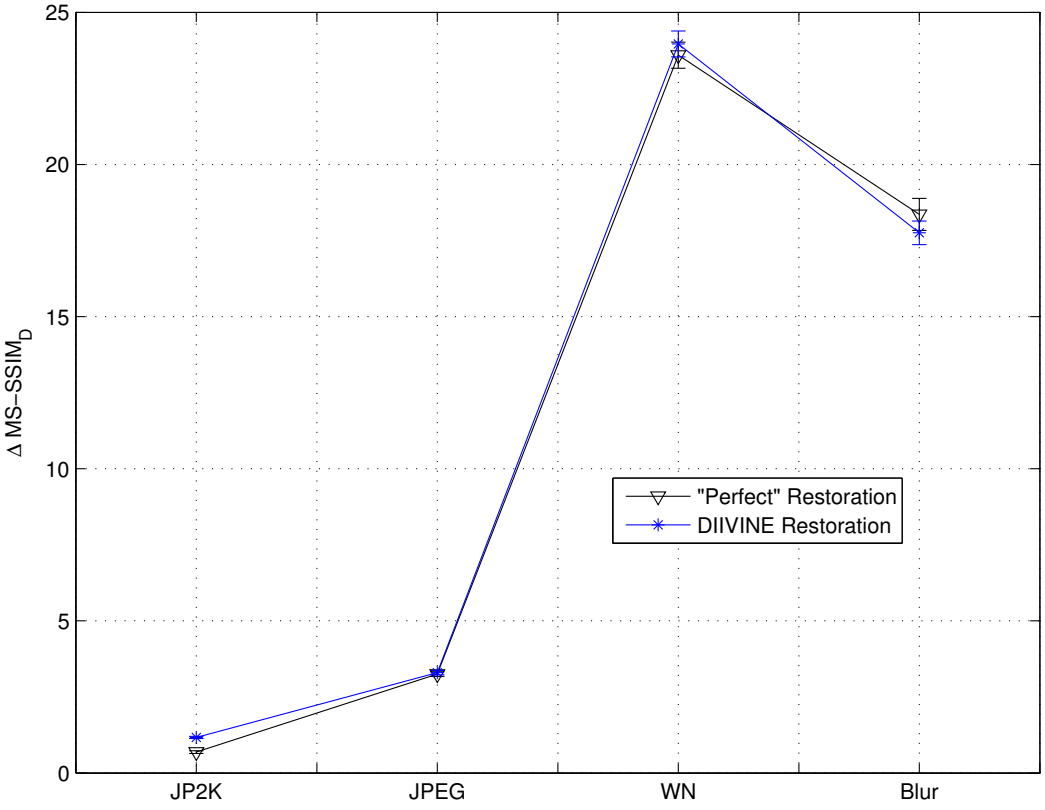


Figure 4.6: Mean increments in quality and the standard error bars for perfect repair, DIIVINE-based and BRISQUE-based GENII-1 algorithms.

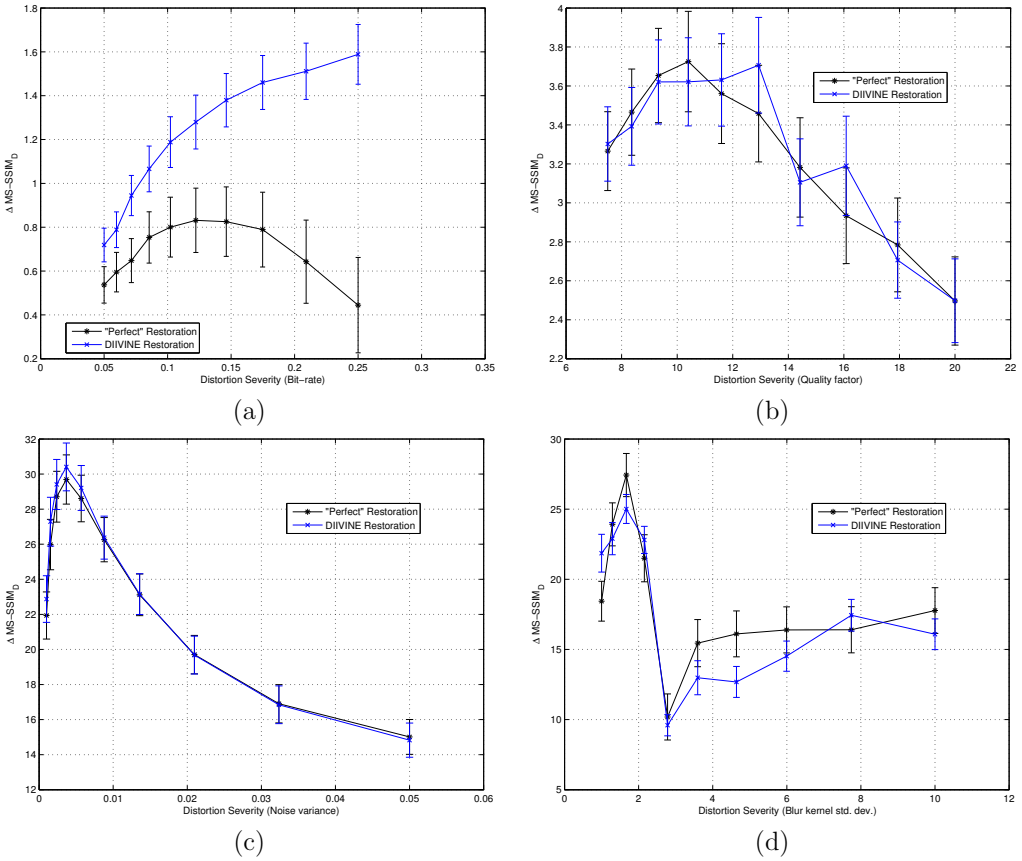
deconvolution and denoising.

In order to study the effect of the amount of distortion on the performance of GENII-1, Fig. 4.7, plots the gain in quality (as measured by MS-SSIM$_D$) as a function of distortion severity for each distortion type. Again, the gains that would be obtained by the (unfair) baseline approach equipped with perfect knowledge of the distortion type and the input distortion parameters is also plotted for comparison purposes. For all distortions, an increase in severity reduces gains in objective quality (MS-SSIM$_D$). The results indicate that the perceptually optimized GENII-1 performs as well as (if not better than) the 'perfect' repair scheme, even though the 'perfect' scheme has full knowledge of the distortion parameters.

The case of JP2K requires further explanation. Although the perfect repair algorithm is a somewhat unfair baseline, it does not have the added advantage of quality-driven self-correction and so, in many cases the output images obtained are of inferior quality relative to the input distorted image.

The training procedure for WN was also modified to provide perceptually optimized denoised images, and the perfect reconstruction baseline does not have this training-based advantage. Although the gains obtained are not reflected in the mean quality-gain plots, on individual images, such a training procedure does indeed produce better quality, as exemplified by Fig. 4.3.

Finally, to provide a visual illustration of the results, Fig. 4.8 plots samples of distorted images from the test set and their repaired versions us-



(a) (b) (c) (d)

Figure 4.7: Mean changes in objective quality (MS-SSIM$_D$) and the standard error bars for perfect repair, DIIVINE-based and BRISQUE-based generalized repair as a function of distortion severity for: (a) Deringing, (b) Deblocking, (c) Denoising and (d) Deblurring.

ing GENII-1 with DIIVINE features along with the quantitative changes in objective quality that were obtained.

4.2.2.4 Iterative Image repair

Iterative repair using GENII-1 can proceed as illustrated in Fig. 4.1, where the repair is performed in a loop until a stopping condition is reached. This condition could be a pre-fixed threshold on quality (which may not always be achieved) or one that assesses the amount of improvement in quality, and stops when the improvement becomes small, ceases to be positive or some combination.

The improvement in quality obtained could be computed relative to the original distorted image at each iteration, or as a difference between the quality at the current iteration and the previous one. Since the repair chain is not guaranteed to produce a steady improvement in quality at each iteration, the best solution would be the former, where one computes the difference between the current quality and the distorted image quality and continues the loop until the improvement is negative. The algorithm would then pick the intermediate image which yields the highest predicted perceptual quality. While this solution is optimal in the current setup, it is time consuming, and stopping if the quality change is negative as compared to the previous iteration may be an attractive alternative in a practical implementation. We now demonstrate two examples of iterative repair using these stopping criteria in Figs. 4.9 and 4.10.

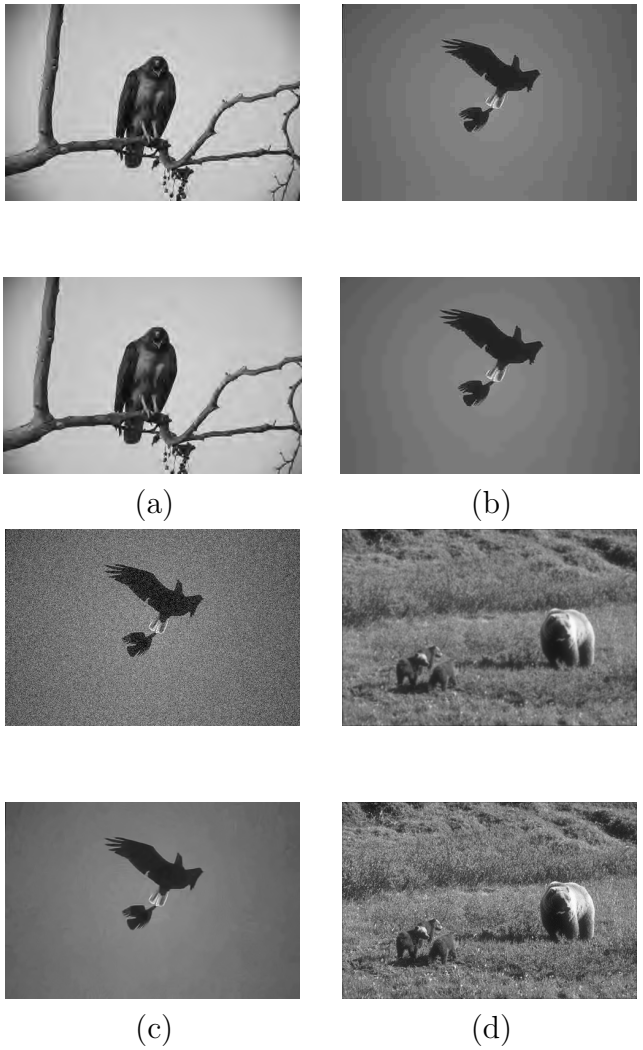


Figure 4.8: Sample distorted images and their repaired versions obtained using the proposed blind general purpose image repair framework. Distortions (Quality Gains): (a) JP2K (7.90), (b) JPEG (15.27), (c) WN (70.60), (d) Blur (51.18).

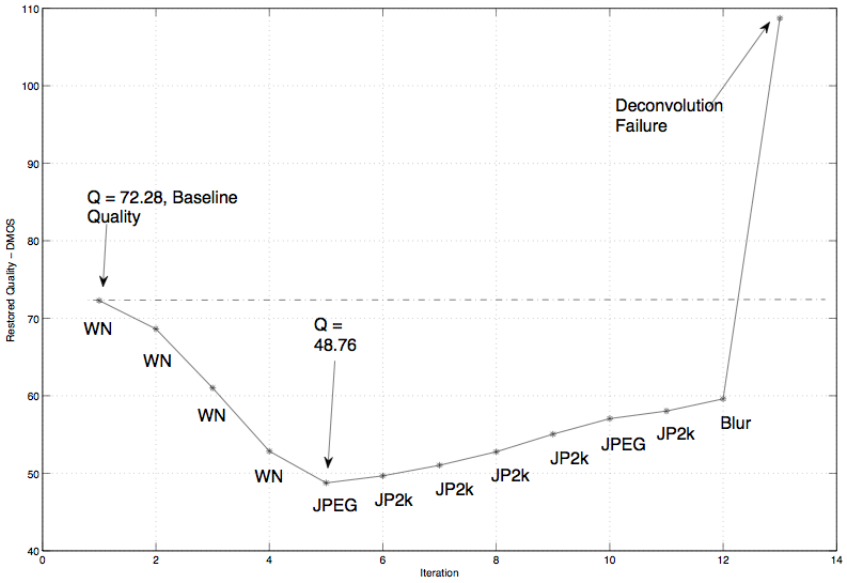
Figure 4.9 illustrates a case where the quality increases steadily with iteration count reaching a well-defined minimum at the fifth repair iteration. Beyond this point, deconvolution failure occurs at iteration 13, and the algorithm stops. Due to our design, the best quality image (iteration 5) is produced as output. In this case, the stopping criterion which checks only the previous iteration would have produced the same result.

Figure 4.10 plots a case where the objective quality degrades in iteration 3 as compared to iteration 2. The simpler stopping criterion would have produced as output the image at iteration 2 (Fig. 4.10 (b)). However, the quality score at iteration 4 is far lower (better) than at iterations 2 and 3. If the more exhaustive stopping criterion were used, the image at iteration 4 would be produced as the output. Deconvolution failure occurs at iteration 6 (image not shown), where the algorithm stops.

4.3 Discussion and Conclusion

We have introduced a distortion-blind perceptually optimizable general-purpose image repair paradigm called GENII that repairs images distorted by any of multiple distortions by using natural scene statistics to (1) identify the likely distortion(s) impairing the image, (2) to estimate the quality of the distorted image, (3) and to estimate the parameters (i.e., distortion severity) of the distortion, and (4) based on these estimated data, selects an appropriate (possibly non-blind) repair module. Steps (1)-(3) are performed using NSS features extracted from the image in the wavelet domain [179]. We demon-

(a) (b) (c)



(d)

Figure 4.9: Example iterative image repair using GENII-1 driven by DIIVINE features, see text for explanation. (a) Distorted image, (b) Best quality repaired image, (c) Deconvolution failure at iteration 13, (d) Quality as a function of repair iterations with predicted distortion type labels. GENII-1outputs (b).

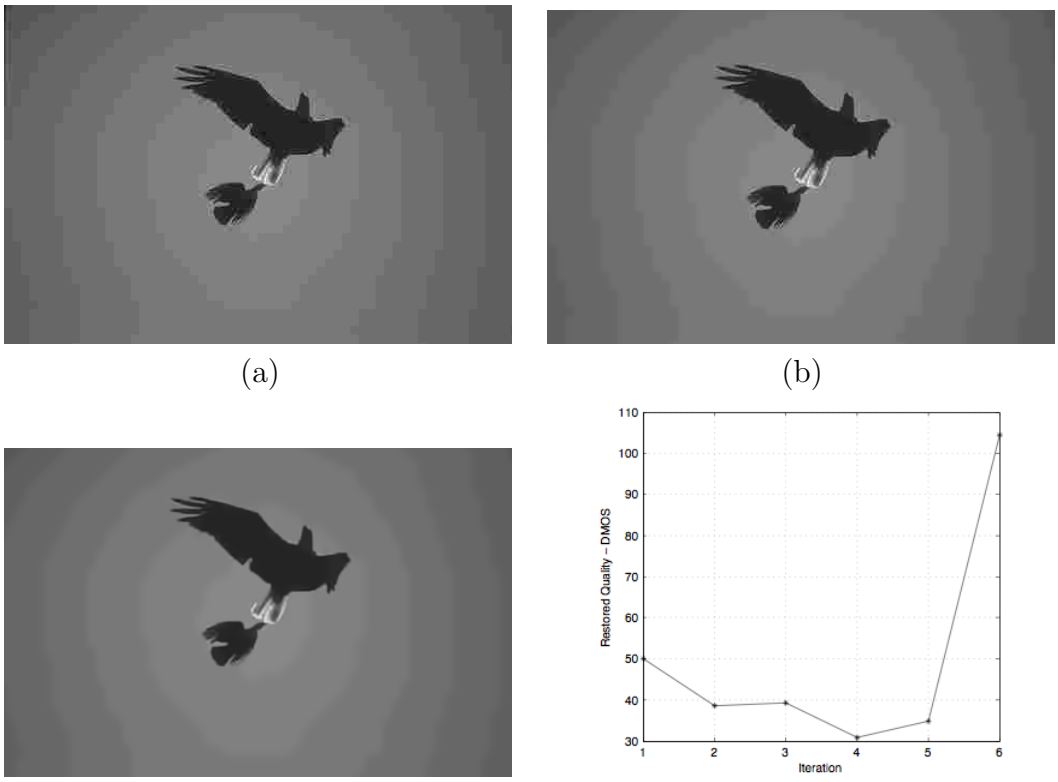


Figure 4.10: Example iterative image repair using GENII-1 driven by DIIVINE features, see text for explanation. (a) Distorted image (MS-SSIM$_D$ = 50.1), (b) repaired image at iteration 2 (MS-SSIM$_D$ = 38.51) , (c) repaired image at iteration 4, highest quality (MS-SSIM$_D$ = 30.81), (d) Quality as a function of repair iterations. GENII-1 outputs (c).

strated a working prototype, dubbed GENII-1, capable of repairing images impaired by any of JPEG200, JPEG, additive Gaussian white noise or linear Gaussian blur distortions.

GENII is the first system of its kind that first identifies any of multiple unknown distortions coming from a trained set and then attempts to fix the image. It is modular and easily extensible to distortions beyond those considered here. The modularity of the approach implies that one could replace any of the repair modules used (e.g., by GENII-1) with better or more appropriate repair algorithms leading to even better performance. New modules could be added for additional distortion classes and/or types, including multi-distortions. This will require design and creation of suitable distorted image databases and associated human studies. Further, the GENII framework was designed such that the output image quality will always be at least as good as the input image quality, as measured by a high-quality, objective but perceptually relevant image quality assessment algorithm, thereby accounting for failures of the repair modules. The iterative nature of GENII implies that distortions introduced by the repair algorithms may also be eliminated, thereby increasing the quality of the output image.

The GENII framework is a radically different approach to image repair, that seeks to maximize the visual quality of the images, as measured by a no-reference image quality assessment algorithm, instead of simply targeting the distortion(s) present in the image. We showed that this new approach to image repair leads to significant improvements in output quality both visually and

quantitatively as measured by a high-performance full-reference objective image quality assessment algorithm. To the best of our knowledge, the proposed framework is the first of its kind to approach the general image repair problem from a perceptual optimization point of view although related problems such as objective quality-driven models for image restoration [43], denoising [44], compression [45, 110], and deblocking [337] have been studied. It is also the first model to combine a no-reference image quality index with distortion identification to perform general purpose image repair.

Our current and future work involves extending the GENII concept to distortions even more diverse and extensive than those considered here and modeling images distorted by multiple coincident distortion and suitable repair processes for such images. The development of GENII models capable of handling multiple distortions will require a number of new developments. First, since the GENII framework is NSS-based, studies of NSS of multiply distorted images will need to be undertaken. This will necessarily include studies of human subjective judgements of these multi-distortions, since the complex way in which the distortion may interact, both in terms of the way in which they modify image structure and the way they are perceived, are likely to be nonlinear and complicated. In essence, they must be viewed as new distortions. Such a deeper NSS and perception-based analysis will enable modifications of current no-reference IQA indices, such as DIIVINE, to handle such multi-distortions.

Chapter 5

Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies

Global mobile data traffic nearly tripled in 2010 for the third consecutive year, exceeding three times the data volume of the entire global Internet traffic just 10 years ago [57]. According to the Cisco Visual Networking Index (VNI) global mobile data traffic forecast, mobile video traffic accounts for nearly 50% of mobile traffic, and it is predicted that this percentage will steadily increase to more than 75% by 2015. As smartphone usage explodes along with mobile enabled video streaming websites such as Amazon Video on Demand, Hulu, Itunes, Netflix and YouTube[1], it is clear that video traffic on mobile devices will continue to account for an increasingly significant portion of mobile data traffic. While this bodes well for end-users able to watch HD quality video clips at the touch of a button, the picture is not completely rosy for those who provide the spectrum.

In early 2010 U.S. Federal Communications Commission (FCC) Chairman Julius Genchowski summarized the problem succinctly - "The record is

[1]Netflix usage accounts for almost 30% of all downstream traffic during peak hours; YouTube accounts for just over 11% (as of May 2011) [241].

pretty clear that we need to find mode spectrum" [215]. According to Peter Rysavy, a wireless analyst, mobile broadband will surpass the spectrum available in mid-2013 [107]. The paucity of bandwidth is evident from the bandwidth caps that most of the wireless providers in the U.S. have recently imposed on data-hungry users.

Given that video traffic accounts for a significant portion of this mobile data traffic, the development of frameworks for wireless networks is a topic of intense study. One particularly promising direction of research is *perceptual optimization* of wireless video networks, wherein network resource allocation protocols are designed to provide video experiences that are measurably improved under perceptual models.

The final receivers of most videos transported over wireless networks are humans and therefore visual perception is the ultimate arbiter of the received visual experience. The human visual system (HVS) is complex and highly non-linear, so treating video data as any other data in solving the resource allocation problem can lead to suboptimal end-user perceptual experiences. The study of models for resource optimization that model video traffic using perceptually relevant features is easily motivated. A key ingredient in developing these tools is understanding and predicting user perception of video quality on mobile devices by conducting large scale human/subjective studies.

Almost all of the studies described in Chapter 2 suffer from several of the following problems : (1) the dataset is of insignificant size, (2) the distortions and their severities considered are insufficient to make judgments

on perception of quality, (3) the videos were obtained from unknown sources and contain unknown corruptions, (4) the video resolutions are too small to be relevant in today's world, (5) the human studies were conducted on a single device with a fixed display resolution and (6) the database is not publicly available. Realizing the need for an adequate and more modern resource, we have endeavored to create a database of broad utility for modeling and analyzing contemporary wireless video networks. The database enables a new avenue of research – behavioral modeling of visual quality perception. The database and the subjective opinion scores (including the temporal scores) are being made available online in order to help further research in the area of visual quality assessment.

Here, we describe an extensive study that we have recently conducted in order to gauge subjective opinion on HD videos when displayed on mobile devices.

5.1 Subjective Assessment of Mobile Video Quality

5.1.1 Source Videos

The source videos were obtained using a RED ONE digital cinematographic camera. The sequences of REDCODE (.r3d) images received from the MYSTERIUM sensor, using the RED $50 - 150$ mm and $18 - 50$ mm T3 zoom lens were stored as 12-bit REDCODE RAW data, at a resolution of $2K(2048 \times 1152)$ at frame rates of 30 fps and 60 fps using the REDCODE 42MB/s option to ensure the best possible acquisition quality. A tripod was

used in most scenes and the ISO was set in the range 100 to 360 according to the weather – ISOs of 100 or 200 were used for outdoor scenes and 200 or 360 were used for indoor scenes; the shutter speed varied between 1/48 to 1/60 s. The automatic white balance mode was used. The RED drive was used to record the videos.

The source videos were then downsampled to resolution 720p (1280×720) and frame-rate of 30 fps, and the .r3d videos were converted into uncompressed .yuv files using a combination of the `imresize (option : bicubic)` function in MATLAB and VirtualDub. All of the source videos in the database are of duration 15 seconds. A total of 12 videos were selected for this study from a larger subset. These were chosen to be representative of a wide variety of content types that the user might experience. Two of these videos were used to train the subjects (see below) while the rest of the videos were used to perform the actual study. The list below describes each of the videos used in the study.

1. Friend Drinking Coke (fc) : Shot at studio with tungsten light and gel. It shows different light ratios on the face with detailed muscle changes occurring under dim lighting. The camera was fixed.

2. Two Swan Dunking (sd): Shot at Lady Bird Lake, Austin Texas on a sunny morning. There are bright twinkles on the waves, and swans are seen dunking into the water. The camera tracked two of the swans.

3. Runners Skinny Guy (rb) : Shot at a marathon race early in the morning. Many runners show diverse contrasts and colors and complex motions. The fixed camera zooms in and out.

4. Students Looming Across Street (ss) : Shot on the campus of The University of Texas at Austin on a windy morning. Walking students loom towards the camera.

5. Bulldozer With Fence (bf) : Shot at a construction area on a sunny afternoon. Different exposures of light, shadowing of trees, motion of bulldozer and complex textures produce a variegated scene. The camera pans across the screen from left to right.

6. Panning Under Oak (po): Shot under a large oak tree under a blue sky on a sunny afternoon. Many small leaves are visible moving slowly.

7. Landing Airplane (la) : Shot at Austin-Bergstrom International Airport on a cloudy afternoon. The landing airplane exhibits fast motion, and the background changes rapidly. The camera tracked the airplane from upper right to lower left.

8. Barton Springs Pool Diving (dv) : Shot at Austin's Barton Springs Pool on a sunny afternoon. There are sparsely moving people, and one diver who creates a splash. The camera was fixed.

9. Trail Pink Kid (tk) : Shot at a Lady Bird Lake trail on a sunny morning. People walk or jog at various speeds in different directions. The camera

was fixed.

10. Harmonicat (hc) : Shot at Zilker Park in Austin on a sunny afternoon. A musician plays guitar and harmonica in front of a tree. The camera zooms in and out.

11. Fountain Vertical (fv) : Shot at LBJ Library fountain on the campus of The University of Texas at Austin on a sunny morning. The fountain jets water into the air in front of a campus skyline. The camera was fixed.

12. Hyein BSP (hy): Shot at Austin's Barton Springs Pool on a sunny afternoon. A child with a colorful dress walks next to the water. The camera pans the scene from right to left.

Figure 5.1 shows sample frames from the various video sequences.

5.1.2 Distortion Simulation

Each of the reference videos were subjected to a variety of distortions including: (a) compression, (b) wireless channel packet-loss, (c) frame-freezes, (d) rate adaptation and (e) temporal dynamics. In this section we detail how these distorted videos were created.

5.1.2.1 Compression

We used the JM reference implementation of the H.264 scalable video codec (SVC) to compress the 720p HD reference videos [116, 117, 244]. Since

$(a)fc$ $(b)sd$ $(c)rb$

$(d)ss$ $(e)bf$ $(f)po$

$(g)la$ $(h)dv$ $(i)tk$

$(j)hc$ $(k)fv$ $(l)hy$

Figure 5.1: Example frames of the videos used in the study. fv and hy were used for training the subjects while the rest of the videos were used in the actual study.

the SVC implementation does not allow rate control for layers above the base layer, we use fixed QP encoding. The QP was varied across videos and layers in order to produce the target bit-rates for each layer of every video. The videos were compressed using 6 SNR layers (temporal and spatial scalability were not evaluated in this study), and 4 of these layers (R_1, R_2, R_3, R_4; $R_1 < R_2 < R_3 < R_4$) were manually chosen for each video based on their perceptual separation. As other authors have argued, ensuring perceptual separation between the videos in QA studies makes it possible for humans (and algorithms alike) to produce consistent judgements of visual quality [182, 255].

Since the video content is quite varied, the bit-rates for each of these layers varies across videos; all videos were compressed with rates between 0.7 Mbps and 6 Mbps. The choices of rates were based on commonly-used parameters for transmission of HD videos over networks as well as rates that are generally seen on wifi networks. The videos were encoded with an intra period of 16 and loss aware distortion optimization (LARDO) was enabled with packet-loss rates set to 3%. Instead of fixing the number of macroblocks per slice, the number of bytes per packet was fixed at 200 bytes – as recommended for wireless transmission of H.264 coded video [273].

Thus, for each video, four compressed SVC streams were created, yielding a total of 40 compressed videos.

5.1.2.2 Wireless channel packet-loss

H.264 SVC compressed videos were transmitted over a simulated wireless channel in order to induce loss, thereby affecting perceptual quality. The simulated channel was modeled using an IEEE 802.11- based wireless channel simulator implemented in LabVIEW. The system comprised of a single link channel with coding, interleaving, QAM modulation, and OFDM modulation. A bit stream containing 2,000,000 bits was sent through a frequency selective channel with 5 taps at an SNR of 15 dB; 4QAM and a 1/2 rate convolutional code were used. These kinds of a bit-streams were sent 100 times, and for each transmission an error trace was created by XORing the transmitted bit-stream with the received bit-stream, which recorded the erroneous bit-locations. These error traces were used to induce errors in the compressed video streams. For each video, a random error-trace from the set of 100 traces was picked and applied, where a video packet was considered to be lost if one of the bits of the packet was erroneous [273]. Since the SVC decoder imposes certain requirements on decoding the video due to the layered architecture, care was taken to ensure that the loss of packets would not result in an error at the decoder.

Each of the compressed videos was transmitted over the wireless channel, resulting in a total of 40 wireless channel distorted videos.

5.1.2.3 Frame-freezes

Two kinds of frame-freeze models were used to create distorted videos: frame freezes for (1) stored video delivery and (2) live video delivery. In the case of stored videos, frame-freezes do not result in the loss of a video segment from the video, i.e., the videos maintain temporal continuity after the freeze. On the other hand, frame-freezes in live video delivery result in a loss of video segments, i.e., a lack of temporal continuity.

For both of the above cases, the model for frame-freeze is as follows. For every x seconds of freeze (where the last frame in the buffer is displayed on the screen until the next frame arrives), the post-freeze video playback is of duration bx seconds ($b > 1$), i.e., the longer the user waits, the longer the post-freeze playback. In our simulations we chose $b = 1.5$.

Three stored video freeze lengths were modeled: (i) 1 second (short bursts of video playback with 8 freezes), (ii) 2 seconds (longer video playback, with 4 freezes) and (iii) 4 seconds (2 freezes, longest continuous video playback); the live video freeze length was set to be 4 seconds. In all cases, there was a lead-in time of 3 seconds, i.e., the first 3 seconds of the video playback did not incorporate a freeze. All frame-freezes were simulated on uncompressed reference videos.

A total of 30 frame-freeze distorted videos (3 for each reference video) were thus obtained.

5.1.2.4 Rate Adaptation

Psychovisual studies have demonstrated that humans are more sensitive to changes in a visual stimulus than to the magnitude of the stimulus [303]. In order to investigate whether such behavior translates to judgments of temporal quality, we simulated rate-changes as a function of time as the subject views a particular video. Specifically, the subject starts viewing the video at a rate R_X, then after n seconds switches to a higher rate R_Y, then again after n seconds switches back to the original rate R_X. Comparing such a rate-adapted stream with the appropriate compressed stream may provide important information regarding human behavioral responses to time-varying video data rates.

Such a scheme may also reveal whether humans prefer shorter durations of high quality content in the midst of a low quality stream, or if they prefer to view the low quality stream without any fluctuation in quality. Thus we may find answers to questions like: Does exposing the viewer to better quality increase his expectations, thereby reducing his quality rating for the lower quality segment of the stream? From a resource allocation perspective this condition will provide data that will allow for better allocation of resources, where 'better' is a function of the quality perceived by the end user. This condition may provide answers to questions like: Given that the channel is going to allow a rate higher than the current one for only n seconds before one is forced to revert back to the current rate, should one switch to a higher rate for n seconds, given that you are currently at rate R_X?

It should be clear from the above discussion that such behavioral as-

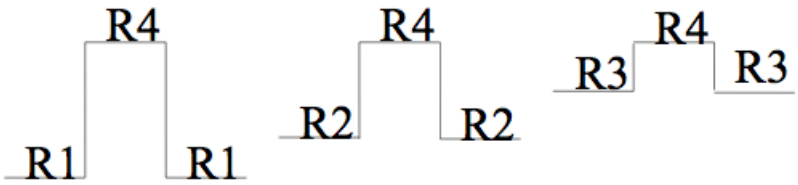


Figure 5.2: Rate Adaptation: Schematic diagram of the three different rate-switches in a video stream simulated in this study.

pects of quality perception may be a function of the difference between the initial rate and the final rate, as well as of the initial rate itself. Hence, we simulate three different rate switches, where $R_X = R_1, R_2$ and R_3 and $R_Y = R_4$. Although the duration n is another potential influence on human behavior, because of on the length of the subject's sessions, we fixed $n = 5$.

The three rate-adaptations which are illustrated in Fig. 5.2 yielded to a total of 30 rate-adapted distorted videos.

5.1.2.5 Temporal Dynamics

In the previous section, we simulated conditions that evaluated the effect that a single rate switch has on perceived quality. One would imagine that the subjective perception of quality is also a function of the *number* and *lengths* of the rate-switches that occur in a stream. In order to evaluate this, we simulated a multiple rate-switch condition, where the rate was varied between R_1 to R_4 multiple times (3). This is illustrated in Fig. 5.3. To ensure an objective comparison between the multiple and single rate-change scenarios, the two conditions are simulated such that the average bit-rate was the same

Figure 5.3: Temporal Dynamics: Schematic illustration of two rate changes across the video; the average rate remains the same in both cases. Left: Multiple changes and Right: Single rate change. Note that we have already simulated the single rate-change condition as illustrated in Fig. 5.2, hence we ensure that the average bit-rate is the same for these two cases.

in both cases.

Apart from multiple switches, one may intuit that subjective quality is also influenced by the *abruptness* of the switch, i.e., instead of switching directly between R_1 and R_4, it may be useful to evaluate conditions where the rate is first switched to an intermediate level Rz from the current level and then to the other extreme. Studying responses to this condition may reveal whether easing a user into a higher/lower quality regime is better than abruptly switching between these two regimes. It should be clear that the intermediate rate Rz may have an impact on the perception of quality as well. Hence, we simulated the following rate-switches: (1) $R_1 - R_2 - R_4$, (2) $R_1 - R_3 - R_4$, (3) $R_4 - R_2 - R_1$ and (4) $R_4 - R_3 - R_1$, as illustrated in Fig. 5.3. Again, the average bit-rate remains the same across these conditions as well as over the conditions in Fig. 5.3.

Notice that the rate-changes illustrated in Fig. 5.4 form dual structures – including such models may also reveal whether the user is influenced by the

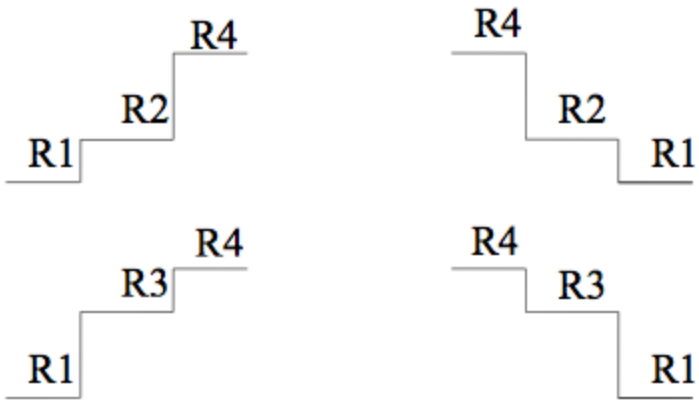


Figure 5.4: Temporal Dynamics: Schematic illustration of rate-changes scenarios. The average rate remains the same in all cases and is the same as in Fig. 5.3. The first row steps to rate R_2 and then steps to a higher/lower rate, while the second row steps to R_3 and then back up/down again

quality observed towards the end of the video. Specifically, we seek to answer the question: Which of the following scenarios is preferable: ending the video with a high quality segment, or ending the video with a low-quality segment? Again,in addition to supplying data on human behavioral responses to time-varying video quality, answering these kinds of questions may also facilitate making better resource allocation decisions. A total of 50 distorted videos with varying temporal dynamics were thus created.

While it is impossible to plot all of the various temporal distortions simulated here, Figs. 5.5 and 5.6, show two examples of distorted frames from the distorted videos, along with the reference frames for comparison. The reader is invited to download the freely available database, in order to better

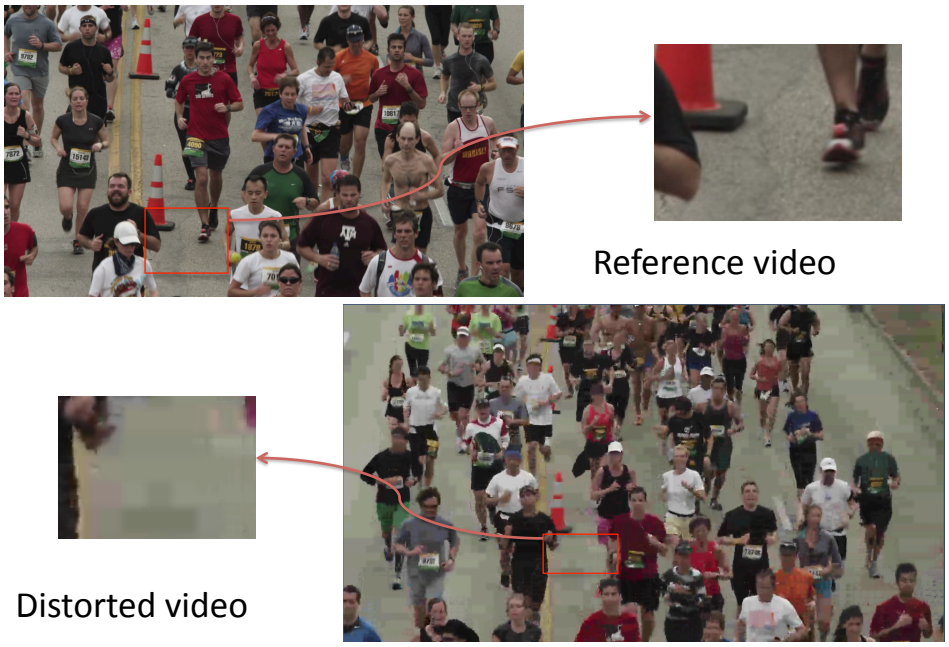


Figure 5.5: Figure illustrating the spatial effect of the distortions simulated in this study for a frame from video 'rb'. Also plotted are the reference frame and a zoomed area for comparison purposes.

visualize the distortions.

In summary, the LIVE Mobile VQA database consists of 10 reference videos and 200 distorted videos (4 compression + 4 wireless packet-loss + 4 frame-freezes + 3 rate-adapted + 5 temporal dynamics per reference), each of resolution 1280×720 at a frame rate of 30 fps, and of duration 15 seconds each.

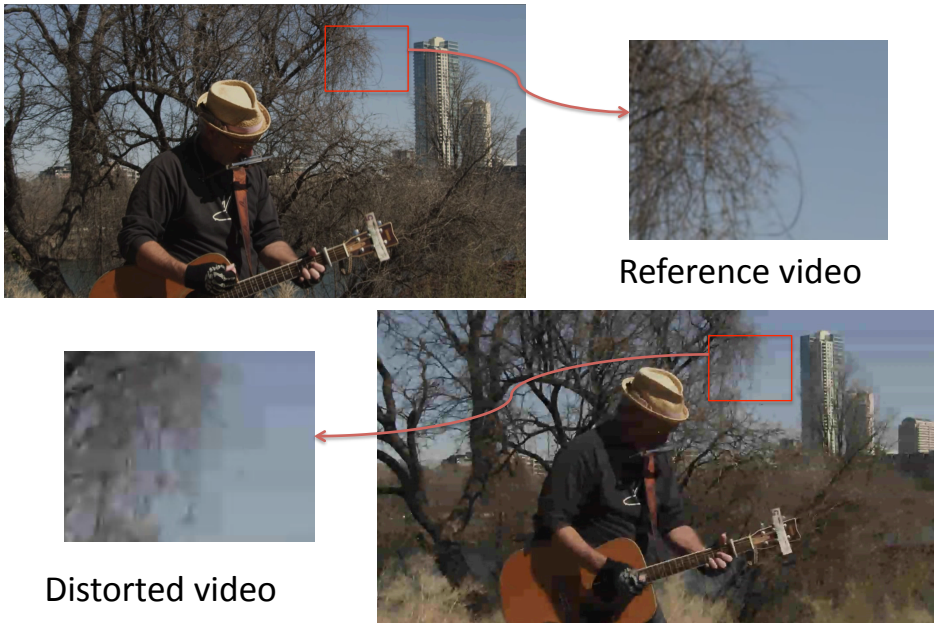


Figure 5.6: Figure illustrating the spatial effect of the distortions simulated in this study for a frame from video 'hc'. Also plotted are the reference frame and a zoomed area for comparison purposes.

5.1.3 Test Methodology

5.1.3.1 Design

A single-stimulus continuous quality evaluation (SSCQE) study [288] with hidden reference [182, 217, 255] was conducted over a period of three weeks at The University of Texas at Austin, LIVE subjective testing lab. Each subject was asked to view and rate the videos one video at a time. Each original, uncompressed reference video was randomly placed amongst the set of videos shown to each user in each session, although the subjects are unaware of their presence. The score that the subjects gave these 'hidden' references is representative of the bias that the subject carries. By subtracting the reference video scores from those for the distorted videos, the biases are compensated for yielding differential scores for each distorted video. We believe that SS with hidden reference studies are preferable to longer double-stimulus (DS) studies [182, 255]. Shorter studies make the the study duration less likely to fatigue the subjects, while allowing the subjects to evaluate a larger set of conditions, for a given study duration. Perhaps most importantly, a SS study design better models real video experiences; typical users deploying mobile video devices in their daily activities are unlikely to ever encounter side-by-side or sequential back-to-back video comparisons. Moreover, unlike a TV showroom, the visual distortions we are interested in are display-device independent and occur in isolation. The choice of a continuous scale as opposed to a discrete 5-point ITU-R Absolute Category Scale (ACR) has advantages: expanded range, finer distinctions between ratings, and demonstrated prior efficacy [182, 255].

5.1.3.2 Display

The user interface was developed on Eclipse[2] using the Android SDK, since the target platforms for the human study were Android-based devices. Although the platform did not allow for explicit control over the video buffer as is allowed by the XGL toolbox [7] which we have previously used [182, 255], no errors such as latencies were encountered while displaying the videos. Since the Android platform does not allow for RAW video playback, the RAW videos were embedded in a 3gp container and compressed using the MPEG-4 codec via ffmpeg. While this additional compression was undesirable, the choice of the platform made this unavoidable. However, the bit-rate for compression was $> 18Mbps$ with the QP set at 0 on ffmpeg, and we were unable to detect any differences between the embedded 3gp streams and the original YUV videos.

The videos were displayed on two devices – the Motorola Atrix smartphone and the Motorola Xoom tablet. The Atrix consists of a dual-core 1 Ghz ARM Cortex-A9 processor, with 1 GB RAM, ULP GeForce GPU and the Tegra 2 chipset. Videos were displayed on the Atrix 4-inch Gorilla glass display with a screen resolution of 960×540; the Atrix is capable of playing out videos at 1080p and the processor was powerful enough to avoid any buffering or playback issues when playing the high-resolution content. The Xoom uses a 1 Ghz NVIDIA Tegra 2 AP20H dual-core processor with 1 GB RAM. Videos were displayed on the 10.1-inch TFT display with a screen resolution

[2]Eclipse is an integrated development environment (IDE) for JAVA, C, C++, Perl amongst other languages, and is freely available: http://www.eclipse.org/.

of 1280×800. As with the Atrix, the Xoom had no problems playing out 720p videos. The devices do not allow for calibration; however, the same devices (with brightness set at max) were used throughout the course of the study.

5.1.3.3 Subjects, Training and Testing

The subjective study was conducted at The University of Texas at Austin (UT) and involved mostly undergraduate students, with a male majority. The study was voluntary and no monetary compensation was provided to the participants. The average subject age was between 22-28 years and the subjects were inexperienced with video quality assessment, types of video distortion and concepts underlying the perception of quality. Though no vision test was performed, a verbal confirmation of soundness of (corrected) vision was obtained from the subject. This approach follows our continuing philosophy towards conducting large-scale image and video quality subjective studies: rigorous visual screening of subjects, such as we routinely do in our other vision science work, may bias results as compared to a 'typical user'. While our philosophy in this regard does not necessarily accord with published (and largely outdated) industry standards, we have discussed our view with other vision scientists and received general accord. We further believe that this approach allows for greater freedom and realism in designing large scale studies such as the one described here, using mobile devices likely to be used in highly diverse conditions and for which there exist no guidelines.

Each subject attended two separate sessions as part of the study such

that each session lasted less than 30 minutes, and the sessions were separated by at least 24 hours, in order to minimize fatigue [288]. Informal after-study feedback indicated that the subjects did not experience any uneasiness or fatigue during the course of the sessions. Each session consisted of the subject viewing 55 videos (50 distorted + 5 reference), and a short training set (6 videos) preceded the actual study. The videos were shown in random order across subjects as well as within a single session for a subject. Care was taken to ensure that two consecutive sequences did not belong to the same reference content, to minimize memory effects [288].

The videos were displayed on the center of the screen with an uncalibrated continuous bar at the bottom, which was controlled using the touchscreen. The subjects were briefed about the bar during the training session. Before the video was played, a screen indicating that the video was ready for playback was displayed. Once the subject hit 'play' the video played on the screen. The subjects were asked to rate the videos as a function of time i.e., provide instantaneous ratings of the videos, as well as to provide an overall rating at the end of each video. At the end of each video a similar continuous bar was displayed on the screen, although it was calibrated as "Bad", "Fair", and "Excellent" by markings, equally spaced across the bar. Although the bar was continuous, the calibrations served to guide the subject. Once the quality was entered, the subject was not allowed to change the score. The quality ratings were in the range 0-5. The instructions to the subject are reproduced in the Appendix.

Fig. 5.7 shows the various stages of the study.

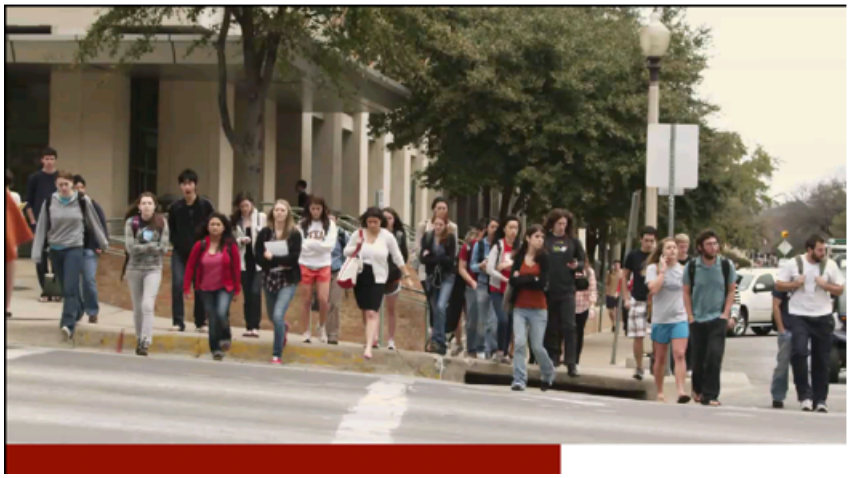
5.1.4 Processing of the Scores

A total of thirty-six subjects participated in the mobile study and seventeen subjects participated in the tablet study. The mobile study was designed so that 18 subjective ratings were obtained for each of the 200 videos in the study. 100 distorted videos from this set of 200 distorted videos were used for the tablet study, and thus each of the 100 videos in the tablet study received ratings from 17 subjects. The subject rejection procedure in [288] was used to reject two subjects from the mobile study, while no subjects were rejected from the tablet study. The scores from the remaining subjects were then averaged to form a Differential Mean Opinion Scores (DMOS) for each video. The DMOS is representative of the perceived quality of the video. Specifically, let s_{ijk} denote the score assigned by subject i to the distorted video j in session k, s_{ijref_k} the score assigned by subject i to the *reference* video associated with the distorted video j in session k, M_j the total number of rating received for video j and let N_{ik} be the number of test videos seen by subject i in session k. The difference scores d_{ijk} are computed as

$$d_{ijk} = s_{ijk} - s_{ijref_k}$$

The DMOS (after subject rejection) is then

$$DMOS_j = \frac{1}{M_j} \sum_i \sum_k d_{ijk}$$



(a)

(b)

Figure 5.7: Study Setup: (a) The video is shown at the center of the screen and an (uncalibrated) bar at the bottom is provided to rate the videos as a function of time. The rating is controlled using the touchscreen. (b) At the end of the presentation, a similar calibrated bar is shown on the screen so that the subject may rate the overall quality of the video.

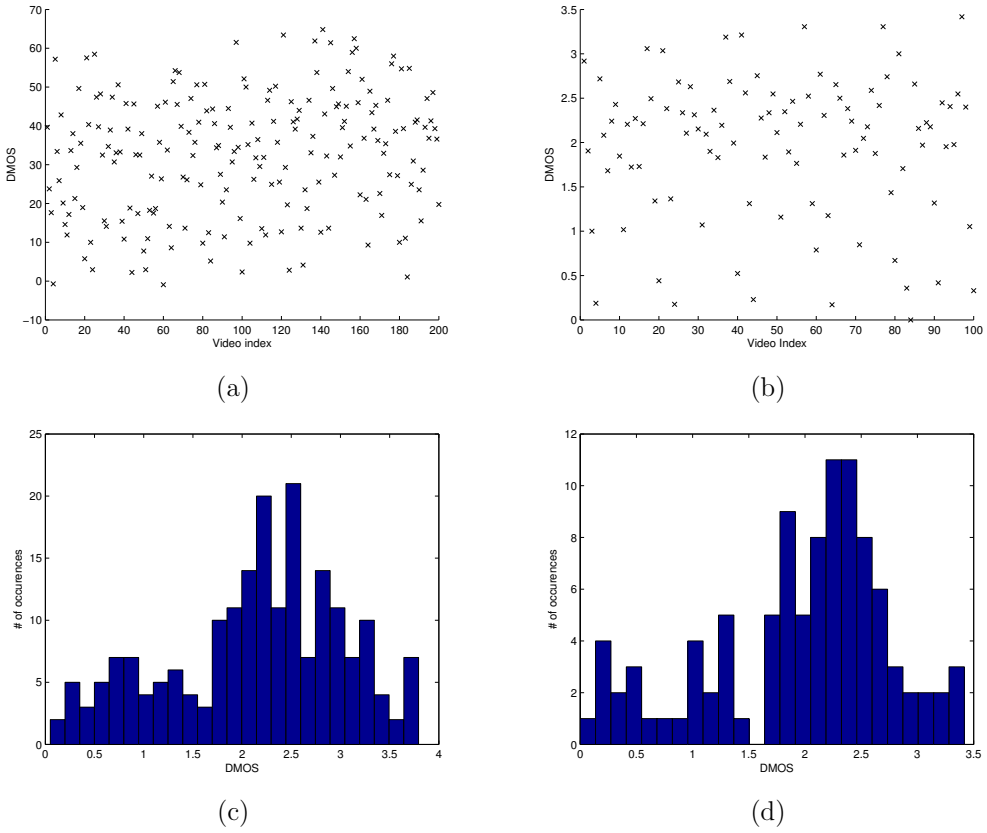


Figure 5.8: DMOS scores for all video sequences: (a) Mobile Study, (b) Tablet Study and the associated histograms of scores for (c) the Mobile Study and (d) the Tablet Study.

DMOS values ideally range continuously from 0 (excellent quality) to 5 (worst quality); however small negative values as possible due to the nature of DMOS computation.

DMOS was computed only for the overall scores that the subject assigned to the videos. Fig. 5.8 plots the DMOS scores across distorted videos for the mobile and tablet studies, and shows the corresponding histograms in

order to demonstrate that the distorted videos span the entire quality range. The average standard error in the DMOS score was 0.2577 across the 200 distorted videos for the mobile study and 0.2461 across the 100 distorted videos for the tablet study. We assume that the DMOS scores sample a Gaussian distribution centered around the DMOS having a standard deviation computed from the differential opinion scores across subjects for all further analysis.

5.1.5 Evaluation of Subjective Opinion

We analyzed the distorted videos with respect to the subjective DMOS for each of the videos and the associated standard deviations of DMOS across the subjects on the mobile and the tablet studies. For each of the subsections below, we conduct a t-test between the Gaussian distributions centered at the DMOS values (and having a associated, known standard deviation) of the conditions we are interested in comparing at the 95% confidence level. Since the conditions being compared are functions of content, we compared each of the 10 reference contents separately for each pair of conditions. In the tables that follow, a value of '1' indicates that the row-condition is statistically superior to the column-condition, while a '0' indicates that the row is worse than a column; a value of '-' indicates that the row and column are statistically indistinguishable from each other. For example, in Table 5.1, for all the 10 contents, videos compressed at rate R_2 have statistically better visual quality than those compressed at rate R_1, while they are statistically worse than those compressed at a rate R_3. Further, for the tablet study, we compared the results

	R_1	R_2	R_3	R_4
R_1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
R_2	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
R_3	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0
R_4	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -

Table 5.1: Mobile Study: Results of t-test between the various compression-rates simulated in the study. A value of '1' indicates that the row is statistically superior (better visual quality) than the column, while a value of '0' indicates that the row is statistically worse (lower visual quality) than the column; a value of '-' indicates that the row and column are statistically equivalent. Each sub-entry in each row/column corresponds to the 10 reference videos in the study.

obtained from the tablet study to those obtained from the mobile study across all distortions as well as for each distortion subsection.

5.1.5.1 Mobile Study

The results from the statistical analysis are tabulated in Tables 5.1 - 5.7. Due to the dense nature of the content, we summarize the results in the following paragraphs. Note that the text only provides a high level description of the results in the table, the reader is advised to thoroughly study the table in order to better understand the results.

Compression (Table 5.1) This table confirms that the distorted videos were perceptually separable. Notice that each compression rate is statistically better (perceptually) than the next lower rate over all content used in the study.

	$F1$	$F2$	$F3$	FR_4
$F1$	- - - - - - - - - -	0 0 0 - 0 0 0 - 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$F2$	1 1 1 - 1 1 1 - 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 - 1 - 1 0 - 0
$F3$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	1 1 1 1 1 1 1 1 1 1
FR_4	1 1 1 1 1 1 1 1 1 1	1 1 1 - 0 - 0 1 - 1	0 0 0 0 0 0 0 0 0 0	- - - - - - - - - -

Table 5.2: Mobile Study: Results of t-test between the frame-freezes simulated in the study. A value of '1' indicates that the row is statistically superior (better visual quality) than the column, while a value of '0' indicates that the row is statistically worse (lower visual quality) than the column; a value of '-' indicates that the row and column are statistically equivalent. Each sub-entry in each row/column corresponds to the 10 reference videos in the study.

Frame-freeze (Table 5.2) For frame-freezes, the following trend is seen across most of the contents: longer freezes are preferred to shorter freezes, which lead to choppy playback, implying playback immediately after the buffer receives data is less desirable than waiting before playback. We also observe that pauses of 4 seconds are seemingly tolerable. For the frame-freezes with lost segments (real-time freezes), one would conjecture that lost segments are important and became evident when the segments are about 4 seconds long or larger. Further, it seems that shorter freezes (choppy playback) are regarded as worse than lost frames.

Rate Adaptation (Tables 5.3, 5.4) While conventional wisdom might dictate that people do not prefer fluctuations in video quality, our study seems to indicate that it is preferable to switch to a higher rate if possible, especially if the duration of the higher rate is at least half the duration of the lower rates. Further, if one is capable of maintaining a continuous rate at a value higher than the base rate of the switch (eg., $R_2 - R_4 - R_2$ vs. R_3), the continuous

	$R_1 - R_4 - R_1$	$R_2 - R_4 - R_2$	$R_3 - R_4 - R_3$
$R_1 - R_4 - R_1$	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$R_2 - R_4 - R_2$	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0
$R_3 - R_4 - R_3$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -

Table 5.3: Mobile Study: Results of t-test between the various rate-adapted distorted videos simulated in the study. A value of '1' indicates that the row is statistically superior (better visual quality) than the column, while a value of '0' indicates that the row is statistically worse (lower visual quality) than the column; a value of '-' indicates that the row and column are statistically equivalent. Each sub-entry in each row/column corresponds to the 10 reference videos in the study.

	R_1	R_2	R_3	R_4
$R_1 - R_4 - R_1$	1 1 1 1 1 1 1 1 1 1	0 0 0 - 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$R_2 - R_4 - R_2$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$R_3 - R_4 - R_3$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 - 0 1 - 0 1 0	0 0 0 0 0 0 0 0 0 0

Table 5.4: Mobile Study: Results of t-test between the various compression-rates and the rate-adapted videos simulated in the study. A value of '1' indicates that the row is statistically superior (better visual quality) than the column, while a value of '0' indicates that the row is statistically worse (lower visual quality) than the column; a value of '-' indicates that the row and column are statistically equivalent. Each sub-entry in each row/column corresponds to the 10 reference videos in the study.

higher rate is preferred.

Temporal Dynamics (Tables 5.5, 5.6) Our analysis indicates that multiple rate switches are preferred over fewer switches, if the subject is able to view the high quality video for longer duration. There is a plausible explanation for this behavior. Our hypothesis is that when shown high quality video for a long time, the bar of expectation is raised, and when the viewer is exposed to low quality segments of the video, s/he assigns a high penalty than on videos

	$R_1 - R_4 - R_1$	$R_1 - R_4 - R_1 - R_4 - R_1$
$R_1 - R_4 - R_1$	- - - - - - - - - -	0 - - - 0 0 0 0 1 -
$R_1 - R_4 - R_1 - R_4 - R_1$	1 - - - 1 1 1 1 0 -	- - - - - - - - - -

Table 5.5: Mobile Study: Results of t-test between multiple rate switches and a single rate switch. A value of '1' indicates that the row is statistically superior (better visual quality) than the column, while a value of '0' indicates that the row is statistically worse (lower visual quality) than the column; a value of '-' indicates that the row and column are statistically equivalent. Each sub-entry in each row/column corresponds to the 10 reference videos in the study.

containing high quality segments of shorter duration. The subject might view the short high quality segments as attempts to improve the viewing experience, thereby boosting overall perception of quality. An even more likely explanation is that long low-quality video segments preceded by much higher quality segments evoke a strong negative response. Of course, our results are conditioned on the degree of quality separation between the low and high quality segments and may not generalize to switches between quality levels exhibiting a lesser degree of quality separation.

Our results also indicate that switching to an intermediate rate before switching to a higher rate is preferred over multiple large-magnitude rate switches, and that the end quality of the video makes a definite impact on perceived quality (see for example, $R_4 - R_3 - R_1$ vs. $R_1 - R_3 - R_4$ in Table 5.6).

Wireless (Table 5.7) The wireless results mirror the compression results, demonstrating the perceptual separability of the videos in the study.

	$R_1 - R_4 - R_1 - R_4 - R_1$	$R_1 - R_2 - R_4$	$R_4 - R_2 - R_1$	$R_1 - R_3 - R_4$	$R_4 - R_3 - R_1$
$R_1 - R_4 - R_1 - R_4 - R_1$	- - - - - - - - - -	- 0 0 0 1 1 0 0 0 0	1 1 - 1 1 1 1 1 1 1	0 0 0 0 - 0 0 0 0 0	1 1 0 0 1 1 1 1 1 1
$R_1 - R_2 - R_4$	- 1 1 1 0 0 1 1 1 1	- - - - - - - - - -	1 1 1 1 1 1 1 1 1 1	0 0 - 0 0 0 - 0 0 0	1 1 1 - 1 1 1 1 1 1
$R_4 - R_3 - R_1$	0 0 - 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	1 - 0 0 0 0 - 0 1 0
$R_1 - R_3 - R_4$	1 1 1 1 - 1 1 1 1 1	1 1 - 1 1 1 - 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	1 1 1 1 1 1 1 1 1 1
$R_4 - R_3 - R_1$	0 0 1 1 0 0 0 0 0 0	0 0 0 - 0 0 0 0 0 0	0 - 1 1 1 1 - 1 0 1	0 0 0 0 0 0 0 0 0 0	- - - - - - - - - -

Table 5.6: Mobile Study: Results of t-test between the various temporal-dynamics distorted videos simulated in the study. A value of '1' indicates that the row is statistically superior (better visual quality) than the column, while a value of '0' indicates that the row is statistically worse (lower visual quality) than the column; a value of '-' indicates that the row and column are statistically equivalent. Each sub-entry in each row/column corresponds to the 10 reference videos in the study.

	WR_1	WR_2	WR_3	WR_4
WR_1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
WR_2	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
WR_3	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0
WR_4	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -

Table 5.7: Mobile Study: Results of t-test between the various wireless packet-losses simulated in the study. A value of '1' indicates that the row is statistically superior (better visual quality) than the column, while a value of '0' indicates that the row is statistically worse (lower visual quality) than the column; a value of '-' indicates that the row and column are statistically equivalent. Each sub-entry in each row/column corresponds to the 10 reference videos in the study.

Comp.	FF	RA	TD	WL	All
0.9493 (0/0.93)	0.7981 (1/0.01)	0.8701 (0/0.56)	0.6298 (0/0.92)	0.9359 (0/0.56)	0.9047 (0/0.89)

Table 5.8: Correlation and results of the Wilcoxon sum-rank test for equal medians (in parenthesis – hypothesis/p-value) between DMOS scores from the mobile and tablet studies. A value of '1' in the brackets indicates that the DMOS scores from the two studies have different medians, while a value of '0' indicates that the medians are statistically indistinguishable at the 95% confidence level.

5.1.5.2 Tablet Study

We compare the results from the tablet study to those from the mobile study for each distortion category and across all the distortions considered here, and tabulate the (linear) correlation coefficient between these two studies in Table 5.8. In the table, we also report the results from a Wilcoxon sum-rank test for equal medians – a value of '1' in the brackets indicates that the DMOS scores from the two studies have different medians, while a value of '0' indicates that the medians are statistically indistinguishable at the 95% confidence level. Also reported are the p-values. The results indicate that while the data is correlated and that the medians are statistically indistinguishable, the degree of correlation is a function of the distortion category. Specifically, for the frame-freeze case, the perception of visual quality varies significantly as a function of the display resolution.

We performed an analysis similar to that for the mobile database and since our results are similar to those for the mobile case, we refrain from reporting those tables here.

5.1.6 Evaluation of Temporal Quality Scores

Recall that we collected subjective opinion scores on time-varying video quality by asking the subject to rate the quality of the video as a function of time. These temporal opinion scores were obtained at a sampling rate equal to that of the frame-rate of the video (i.e., 1/30 fps) for all distortions, except for the frame-freezes where the scores were collected at a rate such that the temporal scores spanned the same support as those for other distortions. Thus a total of 450 temporal scores were collected for each 15 second video. The temporal scores so obtained were then processed as in [253], in order to produce a temporal MOS (z-score) for each video. Specifically, let $f_{ijk}(t)$ be the score assigned to the video j by subject i in session k, where each video is of length T_j. We computed:

$$m_{ik} = \frac{1}{\sum_{j=1}^{N_{ik}} T_j} \sum_{j=1}^{N_{ik}} \sum_{t=1}^{T_j} f_{ijk}(t) \tag{5.1}$$

$$s_{ik}^2 = \frac{1}{\sum_{j=1}^{N_{ik}} T_j - 1} \sum_{j=1}^{N_{ik}} \sum_{t=1}^{T_j} (f_{ijk}(t) - m_{ik})^2 \tag{5.2}$$

$$z_{ij}(t) = \frac{f_{ijk}(t) - m_{ik}}{s_{ik}} \tag{5.3}$$

and finally,

$$MOS_j^f(t) = \frac{1}{M} \sum_{i=1}^{M} z_{ij}(t) \tag{5.4}$$

where $MOS_j^f(t)$ is the mean opinion score recorded over time for video j and M is the number of subjects in the study (after subject rejection, as described earlier).

We analyzed how these temporal scores contribute to the overall perception of visual quality, i.e., how temporal scores might be pooled to reproduce the DMOS that the subject assigned the video at the end of the presentation. The analysis below is simplistic, but much work remains on developing good behavioral models of temporal quality judgements of dynamically changing video distortions. Our first attempt at understanding this new problem is detailed in [253].

We evaluate three different methods of temporal pooling: (1) Mean, (2) Percentile pooling [175, 219, 315], and (3) Memory-effect based pooling.

The temporal mean serves as the baseline and is simply the time-average of $MOS_j^f(t)$. Percentile pooling was proposed in [175, 219, 315] as a method of spatially collapsing image quality scores while emphasizing severe errors. There is some evidence that this type of pooling may relate to the visual quality of videos as well [210]. Here, we sorted the temporal scores in ascending order and averaged the lowest 5% of the sorted scores to produce a single quality score for each video.

One may conjecture that human quality decisions are heavily influenced by the visual quality perceived in the last segment prior to rating. To investigate this claim, we averaged quality scores from a time-window spanning the

	Comp.	FF	RA	TD	WL	All
Mean	-0.9724	-0.2488	-0.9001	0.3374	-0.9729	-0.7008
Percentile Pooling	-0.8970	0.0501	-0.7991	0.0767	-0.9247	-0.7092
Memory Effect ($t = 1s$)	-0.9788	-0.6251	-0.8054	-0.7399	-0.9805	-0.8337
Memory Effect ($t = 2s$)	-0.9777	-0.6309	-0.7861	-0.7082	-0.9794	-0.8360
Memory Effect ($t = 3s$)	-0.9778	-0.6389	-0.7799	-0.6193	-0.9797	-0.8340

Table 5.9: Mobile Study: Correlation coefficient between the temporally pooled subjective scores and the DMOS for various pooling strategies.

	Comp.	FF	RA	TD	WL	All
Mean	-0.9720	-0.1557	-0.9248	0.6757	-0.9847	-0.7031
Percentile Pooling	-0.8543	0.3040	-0.8108	0.4945	-0.9263	-0.5781
Memory Effect ($t = 1s$)	-0.9826	-0.4825	-0.8718	-0.3492	-0.9882	-0.8134
Memory Effect ($t = 2s$)	-0.9850	-0.5565	-0.8343	-0.1702	-0.9899	-0.8092
Memory Effect ($t = 3s$)	-0.9850	-0.5864	-0.8142	0.0794	-0.9900	-0.8116

Table 5.10: Tablet Study: Correlation coefficient between the temporally pooled subjective scores and the DMOS for various pooling strategies.

last n frames of the video, where n is varied between $1 - 3$ seconds in steps of 1 second.

In Tables 5.9 and 5.10, we tabulate the correlation coefficient between the DMOS (as obtained previously) and each of the four pooling strategies, for each distortion as well as across all distortions, for the mobile study and for the tablet study respectively.

Note that the correlations should ideally be negative, since we are comparing the MOS with DMOS; the small positive correlations in the tables are meaningless, and imply that the pooling strategy does not correlate well for those distortion categories.

Tables 5.9 and 5.10 indicate that while the temporal and percentile

pooling strategies are poor approaches to collapsing temporal scores (especially for the frame-freezes and the temporal dynamics case), the memory-effect pooling seems to function better, lending credo to the observation that humans are influenced by the last few seconds of viewing when assessing overall quality. We note that this effect was not observed in the study of [253], but this may have been due to the shorter durations of those videos. We also note that while the Memory-effect does help, the overall improvement achieved is not great, which may be due to the short durations of the clips used in this study. While the videos in this study were at least 50% longer than those in [253, 255], they are still short relative to the kind of Memory effects that can occur.

The tables also indicate that, while most pooling strategies work for videos exhibiting uniform visual quality over time video (for example, compression), almost all pooling strategies performed poorly when the quality changes dynamically – either when the compression rate is varied (eg., temporal dynamics) or if the video freezes. One could conjecture that a good behavioral model of temporal quality pooling should improve correlation with DMOS, and that such temporal pooling models could profitably be incorporated into existing VQA algorithms to provide better predictions of overall visual quality. Finally, we note that temporal pooling had a greater impact in the tablet study than the mobile study. It is possible that the resolution of the display makes dynamically varying distortions even more perceptible on a device with a larger form factor (notice that for compression and wireless

No.	Algorithm
1.	Peak Signal-to-Noise ratio (PSNR)
2.	Structural Similarity Index (SS-SSIM) [311]
3.	Multi-scale Structural Similarity Index (MS-SSIM) [319]
4.	Visual Signal-to-Noise ratio (VSNR) [41]
5.	Visual Information Fidelity (VIF) [261]
6.	Universal Quality Index (UQI) [308]
7.	Noise Quality Measure (NQM) [65]
8.	Signal-to-Noise ratio (SNR)
9.	Weighted Signal-to-Noise ratio (WSNR) [159]

Table 5.11: List of FR 2D IQA algorithms evaluated in this study.

distortions the correlations are similar to those for the mobile study). The results seem to indicate that temporal pooling strategy should account for display resolution as well.

5.2 Evaluation of Algorithm Performance

We evaluated a wide variety of full-reference (FR) IQA algorithms against the human subjective scores collected. Table 5.11 lists these algorithms, all of which are available as part of the Metrix Mux toolbox [92]. The reader is referred to the citations for details on these approaches.

The FR IQA algorithms were applied on a frame-by-frame basis and the average score across time used as a final measure of quality. Since it is unclear how FR QA algorithms may be used for frame-freezes (an interesting and important problem for the future), we did not include this case in our evaluation below.

We also evaluated two FR VQA algorithms – Visual Quality Metric (VQM) [219] and the MOtion-based Video Integrity Evaluation (MOVIE) index [252]. VQM was obtained from [3] while MOVIE is freely available at [248]. The version of VQM that we used (CVQM v13) requires input videos in YUV422p format encased in an avi container. The YUV420p videos were converted to YUV422p using ffmpeg, then placed in an avi container (no compression was used). These algorithms were also not evaluated for their performance on frame-freezes.

5.2.1 Algorithm Correlations Against Subjective Opinion

Tables 5.12 and 5.13, tabulate the Spearman's rank ordered correlation cofficient (SROCC) between the algorithm scores and DMOS for the mobile and tablet studies, Tables 5.14 and 5.15 tabulate the Pearson's (linear) correlation coefficient (LCC) and Tables 5.16 and 5.17, tabulate the root mean-squared-error (RMSE) between the algorithm scores (after non-linear regression, as prescribed in [264][3]) and DMOS.

There are two immediate takeaways from the combined tables. First, that multiscale matters as the display size is reduced. Indeed, the two true wavelet decomposition based algorithms – VSNR and VIF – yielded the best overall performance, exceeding that of true video QA algorithms – the single-scale VQM and the MOVIE index, which is partially-multiscale but omits high

[3]Except for MOVIE, where the fitting failed; instead the logistic specified in [297] was used.

	Comp.	RA	TD	WL	All
PSNR	0.8185	0.5981	0.3717	0.7925	0.6780
SS-SSIM	0.7092	0.6303	0.3429	0.7246	0.6498
MS-SSIM	0.8044	**0.7378**	**0.3974**	0.8128	0.7425
VSNR	**0.8739**	0.6735	0.3170	0.8559	**0.7517**
VIF	0.8613	0.6388	0.1242	0.8739	0.7439
UQI	0.5621	0.4299	0.0296	0.5756	0.4894
NQM	0.8499	0.6775	0.2383	**0.8985**	0.7493
WSNR	0.7817	0.5598	0.0942	0.7510	0.6267
SNR	0.7073	0.5565	0.2029	0.6959	0.5836
VQM	0.7717	0.6475	0.3860	0.7758	0.6945
MOVIE	0.7738	0.7198	0.1578	0.6508	0.6420

Table 5.12: Mobile Study: Spearman's Rank ordered correlation coefficient (SROCC) between the algorithm scores and the DMOS for various IQA/VQA algorithms.

	Comp.	RA	TD	WL	All
PSNR	0.7910	0.4464	0.0981	0.7564	0.5886
SS-SSIM	0.4947	0.3679	0.0773	0.5609	0.4300
MS-SSIM	0.6602	0.4821	0.1400	0.6451	0.5678
VSNR	0.7714	0.4429	0.0469	0.7053	0.5929
VIF	**0.8917**	0.6714	0.0700	**0.8617**	**0.7261**
UQI	0.5053	0.3500	0.0481	0.4226	0.3642
NQM	0.8406	0.4643	0.0792	0.8075	0.6614
WSNR	0.8361	0.6214	0.1462	0.7353	0.6255
SNR	0.7098	0.6321	**0.2354**	0.6602	0.5474
VQM	0.6316	0.4357	0.0515	0.6692	0.5552
MOVIE	0.7744	**0.7714**	0.0658	0.8451	0.6792

Table 5.13: Tablet Study: Spearman's rank ordered correlation coefficient (SROCC) between the algorithm scores and the DMOS for various IQA/VQA algorithms.

	Comp.	RA	TD	WL	All
PSNR	0.7841	0.5364	0.4166	0.7617	0.6909
SS-SSIM	0.7475	0.6120	0.3924	0.7307	0.6637
MS-SSIM	0.7664	0.7089	0.4068	0.7706	0.7077
VSNR	0.8489	0.6581	**0.4269**	0.8493	0.7592
VIF	**0.8826**	0.6643	0.1046	**0.8979**	**0.7870**
UQI	0.5794	0.2929	0.2546	0.7412	0.6619
NQM	0.8318	0.6772	0.3646	0.8738	0.7622
WSNR	0.7558	0.5365	0.0451	0.7276	0.6320
SNR	0.6501	0.3988	0.0839	0.6052	0.5189
VQM	0.7816	0.5910	0.4066	0.7909	0.7023
MOVIE	0.8103	**0.6811**	0.2436	0.7266	0.7157

Table 5.14: Mobile Study: Linear (Pearson's) correlation coefficient (LCC) between the algorithm scores and the DMOS for various IQA/VQA algorithms.

	Comp.	RA	TD	WL	All
PSNR	0.7712	0.4368	0.2520	0.7320	0.6348
SS-SSIM	0.5857	0.4222	0.0814	0.5900	0.4893
MS-SSIM	0.7018	0.5644	0.2134	0.7060	0.6213
VSNR	0.7751	0.5083	0.2202	0.7310	0.6444
VIF	**0.8511**	0.5942	0.0484	0.8541	0.7635
UQI	0.4160	0.2454	0.3043	0.5708	0.3256
NQM	0.8115	0.4124	0.1199	0.8298	0.7178
WSNR	0.8150	0.6704	0.2154	0.7252	0.6665
SNR	0.7158	0.6006	**0.3501**	0.6137	0.5544
VQM	0.6430	0.4897	0.2738	0.7349	0.6150
MOVIE	0.8275	**0.8023**	0.0711	**0.8767**	**0.7828**

Table 5.15: Tablet Study: Linear (Pearson's) correlation coefficient (LCC) between the algorithm scores and the DMOS for various IQA/VQA algorithms.

	Comp.	RA	TD	WL	All
PSNR	0.7069	0.5733	0.4179	0.7279	0.6670
SS-SSIM	0.7566	0.6023	0.4228	0.7670	0.6901
MS-SSIM	0.7316	0.4792	0.4199	0.7160	0.6518
VSNR	0.6021	0.5115	0.4157	0.5932	0.6005
VIF	0.5354	0.5078	0.4572	0.4945	0.5692
UQI	0.9283	0.6496	0.4445	0.7542	0.6916
NQM	0.6374	0.4999	0.4280	0.5463	0.5972
WSNR	0.7458	0.5733	0.4592	0.7707	0.7150
SNR	0.8654	0.6230	0.4580	0.8944	0.7887
VQM	0.7312	0.4840	0.4141	0.7279	0.6663
MOVIE	0.6674	0.4974	0.4458	0.7719	0.6444

Table 5.16: Mobile Study: Root mean-squared-error (RMSE) between the algorithm scores and the DMOS for various IQA/VQA algorithms.

	Comp.	RA	TD	WL	All
PSNR	0.7057	0.5810	0.2510	0.7205	0.6630
SS-SSIM	0.8985	0.5855	0.2585	0.8538	0.7483
MS-SSIM	0.7896	0.5332	0.2533	0.7489	0.6724
VSNR	0.7004	0.5562	0.2530	0.7216	0.6562
VIF	0.5820	0.5195	0.2590	0.5500	0.5541
UQI	1.0080	0.6261	0.2470	0.8683	0.8113
NQM	0.6477	0.5884	0.2575	0.5902	0.5974
WSNR	0.6424	0.4792	0.2532	0.7281	0.6397
SNR	0.7741	0.5164	0.2429	0.8349	0.7141
VQM	0.8047	0.5922	0.2593	0.7594	0.6980
MOVIE	0.6224	0.3855	0.2593	0.5087	0.5342

Table 5.17: Tablet Study: Root mean-squared-error (RMSE) between the algorithm scores and the DMOS for various IQA/VQA algorithms.

frequencies. Multiscale SSIM also does quite well, although it overweights mid-band frequencies. A lesson here is that true multiscale is advisable to achieve scalability against variations in display size, resolution and viewing distance, suggesting future refinements of VQA algorithms.

Secondly as Table 5.12 shows, almost all algorithms fail to reliably predict overall subjective judgements of dynamic distortions – on the set of "temporal-dynamics" distorted videos and to some extent, the set of "rate-adaptation" videos. Some algorithms such as VQM, NQM and VIF perform reasonably well on the wireless distorted videos. For the rate-adaptation case, MS-SSIM and MOVIE were the top performers; however, there clearly remains significant room for improvement. Overall, VSNR, VIF, MS-SSIM and NQM are seemingly well correlated with human perception, while the single-scale UQI is the weakest of the lot probably since it captures the narrowest range of frequencies. The widely criticized PSNR holds its own against compression and wireless distortions, since, while it is not multiscale, it captures high frequency distortions.

The results of algorithms against subjective judgments of videos viewed on the tablet show some interesting contrasts(Table 5.13). Whilst VSNR was the top performer for compression in the mobile case, it does not do as well for the tablet case, where multiscale is less of a factor (at finer scales), with MOVIE and NQM eclipsing it and VIF the clear top performer. Since VSNR is a human visual system (HVS)-based measure which takes the number of pixels per visual degree into account, one could conjecture that a recalibra-

tion of VSNR based on the viewing distance and form factor of the tablet might boost performance. While all the algorithms still have trouble predicting judgments of dynamic distortions, MOVIE successfully predicts judgements of rate-adaptation. On wireless distortions, VIF again does well, as does MOVIE, while VSNR again sees a drop in performance. The performance increase of MOVIE in the tablet wireless case over the mobile case is instructive. Since MOVIE is only partially multiscale and has only been tested against human judgments of videos viewed on larger screens than mobile phones, it is not surprising that its performance improves on videos displayed on screens with a larger form factor. As in the case of VSNR, a recalibration of MOVIE as a function of the form factor, or by making it fully multiscale, would likely improve its performance on smaller screen sizes. PSNR is again close to the end of the pack, with the single-scale UQI being the worst performer.

5.2.2 Hypothesis Testing and Statistical Analysis

5.2.2.1 Inter-algorithm comparisons

We performed a statistical analysis of the algorithm scores in order to gauge if the correlations tabulated above were significantly different from each other. In order to evaluate this, we use the method of [255, 264], where the F-statistic is used to evaluate the difference between the variances of the residuals produced after a non-linear mapping between the two algorithms being compared. We perform a similar statistical analysis and report the results in Figs. 5.9 and 5.10 for the mobile and the tablet studies respectively.

A value of '1' in the figures indicates that the row (algorithm) is statistically better than the column (algorithm), while a value of '0' indicates that the row is worse than the column; a value of '-' indicates that the row and column are statistically identical. In Figs. 5.9 and 5.10, we evaluate this hypothesis for each distortion category as well as for all distortions considered together.

Tables 5.9 and 5.10 validate our observations from the correlations – NQM, VIF, VQM perform well, although interestingly, NQM is the only algorithm that is statistically superior to PSNR overall for the mobile study, while VIF is superior to PSNR in the tablet study, where MOVIE also performed well.

5.2.2.2 Comparison with the theoretical null model

We also performed an analysis to evaluate whether algorithm performances were different from the theoretical null model [255, 264]. Given that we have performed all analysis up to this point using DMOS scores from the database, and given that humans exhibit inter-subject variability , it is important not to penalize an algorithm if the differences between the algorithm scores and DMOS can be explained by the differences between the individual subjective scores and the DMOS. This variance between the differential opinion scores (DOS) and the DMOS is used as a measure of the inherent variance of subjective opinion, and we analyze whether the variances of differences between the algorithm scores and DOS are statistically equivalent to that of DOS and DMOS. Our analysis unfolds as in [255]. Specifically, we compute

	PSNR	SS-SSIM	MS-SSIM	VSNR	VIF	UQI	NQM	WSNR	SNR	VQM	MOVIE
PSNR	- - - - -	1 - - 1 1	- 0 - - -	- - - - -	- 1 1 - -	1 1 1 1 1	- - - 0 0	1 1 1 1 1	1 1 1 1 1	1 0 - - -	1 - 1 1 -
SS-SSIM	0 - - 0 0	- - - - -	0 0 - 0 0	0 - - 0 0	0 1 1 0 0	- 1 1 - 1	0 - - 0 0	- 1 1 - -	- - - - 1	- 0 - 0 0	- - 1 - -
MS-SSIM	- 1 - - -	1 1 - 1 1	- - - - -	- 1 - - -	- 1 - - -	1 1 - 1 1	0 - - 0 0	- 1 - - 1	1 1 - 1 1	- - - - -	- - - - 1
VSNR	- - - - -	1 - - 1 1	- 0 - - -	- - - - -	- 1 1 - -	1 1 1 1 1	- - - 0 -	1 1 1 1 1	1 1 1 1 1	1 0 - - -	1 - 1 1 1
VIF	- 0 0 - -	1 0 0 1 1	- 0 - - -	- 0 0 - -	- - - - -	1 - - 1 1	- 0 0 0 0	1 - - 1 1	1 - - 1 1	1 0 0 1 -	1 0 - 1 -
UQI	0 0 0 0 0	- 0 0 - 0	0 0 - 0 0	0 0 0 0 0	0 - - 0 0	- - - - -	0 0 - 0 0	0 - - - 0	- - - - -	0 0 - 0 0	0 0 - - 0
NQM	- - - 1 1	1 - - 1 1	1 - - 1 1	- - - 1 -	- 1 1 1 1	1 1 - 1 1	- - - - -	1 1 - 1 1	1 1 - 1 1	1 0 - 1 1	1 - 1 1 1
WSNR	0 0 0 0 0	- 0 0 - -	- 0 - - 0	0 0 0 0 0	0 - - 0 0	1 - - - 1	0 0 - 0 0	- - - - -	- - - - -	- 0 0 - 0	- 0 - - 0
SNR	0 0 0 0 0	- - - - 0	0 0 - 0 0	0 0 0 0 0	0 - - 0 0	- - - - -	0 0 - 0 0	- - - - -	- - - - -	- 0 - 0 0	- 0 - - 0
VQM	0 1 - - -	- 1 - 1 1	- - - - -	0 1 - - -	0 1 1 0 -	1 1 - 1 1	0 1 - 0 0	- 1 1 - 1	- 1 - 1 1	- - - - -	- - 1 - -
MOVIE	0 - 0 0 -	- - 0 - -	- - - - 0	0 - 0 0 0	0 1 - 0 -	1 1 - - 1	0 - 0 0 0	- 1 - - 1	- 1 - - 1	- - 0 - -	- - - - -

Figure 5.9: Mobile Study: Statistical analysis of algorithm performance. A value of '1' in the tables indicates that the row (algorithm) is statistically better than the column (algorithm), while a value of '0' indicates that the row is worse than the column; a value of '-' indicates that the row and column are statistically identical. Within each entry of the matrix, the first four symbols correspond to the four distortions (ordered as in the text), and the last symbol represents significance across the entire database.

	PSNR	SS-SSIM	MS-SSIM	VSNR	VIF	UQI	NQM	WSNR	SNR	VQM	MOVIE
PSNR	- - - - -	1 - - - 1	1 - - - -	- - - - -	- - - - 0	- - - - 1	- - - - -	- - - - -	- - - - -	1 - - 1 1	- - - 0 -
SS-SSIM	0 - - - 0	- - - - -	- - - - -	0 - - - 0	0 - - 0 0	- - - - -	0 - - 0 0	0 - - - 0	0 0 0 - -	- - - - -	- - - 0 0
MS-SSIM	0 - - - -	- - - - -	- - - - -	- - - - -	0 - - 0 0	- - - - 1	0 - - 0 0	0 - - - 0	- 0 0 - -	- - - - -	- 0 - 0 0
VSNR	- - - - -	1 - - - 1	- - - - -	- - - - -	0 - - 0 0	- - - - 1	- - - 0 -	- - - - -	- 0 - - -	1 - - - 1	- 0 - 0 -
VIF	- - - - 1	1 - - 1 1	1 - - 1 1	1 - - 1 1	- - - - -	1 - - 1 1	- - - - 1	- - - 1 1	1 - 0 1 1	1 - - 1 1	1 - - - -
UQI	- - - - 0	- - - - -	- - - - 0	- - - - 0	0 - - 0 0	- - - - -	0 - - 0 0	0 - - - 0	- 0 - - 0	- - - - -	- 0 - 0 0
NQM	- - - - -	1 - - 1 1	1 - - 1 1	- - - 1 -	- - - - 0	1 - - 1 1	- - - - -	- 0 - 1 -	- 0 - 1 -	1 - - 1 1	- 0 - - -
WSNR	- - - - -	1 - - - 1	1 - - - 1	- - - - -	- - - 0 0	1 - - - 1	- 1 - 0 -	- - - - -	- - - - -	1 - - - 1	- - - 0 -
SNR	- - - - -	1 1 1 - -	- 1 1 - -	- 1 - - -	0 - 1 0 0	- 1 - - 1	- 1 - 0 -	- - - - -	- - - - -	1 1 1 - -	- - 1 0 0
VQM	0 - - 0 0	- - - - -	- - - - -	0 - - - 0	0 - - 0 0	- - - - -	0 - - 0 0	0 - - - 0	0 0 0 - -	- - - - -	- 0 - 0 0
MOVIE	- - - 1 -	- - - 1 1	- 1 - 1 1	- 1 - 1 -	0 - - - -	- 1 - 1 1	- 1 - - -	- - - 1 -	- - 0 1 1	- 1 - 1 1	- - - - -

Figure 5.10: Tablet Study: Statistical analysis of algorithm performance. A value of '1' in the tables indicates that the row (algorithm) is statistically better than the column (algorithm), while a value of '0' indicates that the row is worse than the column; a value of '-' indicates that the row and column are statistically identical.Within each entry of the matrix, the first four symbols correspond to the four distortions (ordered as in the text), and the last symbol represents significance across the entire database.

the ratio between (a) the variances ($\sigma^2_{algorithm}$) of residuals between the differential opinion scores (DOS) and algorithm scores (after non-linear regression) and (b) the variances (σ^2_{null}) of residuals between the differential opinion scores (DOS) and DMOS for each distortion as well as across all distortions. The ratio of two variances $\sigma^2_{algorithm}/\sigma^2_{null}$ is the F-statistic and at the 95% confidence level, for the degrees of freedom exhibited by the numerator and denominator, one can compute the threshold F-ratio. If the computed F-statistic exceeds the threshold F-ratio, then one accepts the null hypothesis – i.e., the algorithm performance is equivalent to the theoretical null model – else, one rejects the null hypothesis. In Tables 5.18 and 5.19 we report the F-statistic for each distortion and for all distortions for each of the algorithms considered here, as well as the threshold F-ratio for the mobile and tablet study respectively. Fields marked in bold indicate acceptance of the null hypothesis. The tables indicate that across distortions, there does not exist a single algorithm that is equivalent to the theoretical null model, except VIF on the wireless distorted videos. Clearly, there remains much work to do on video quality assessment, both on developing fully scalable VQA algorithms and especially towards understanding human reactions to temporal video dynamics and how to model them.

5.3 Discussion and Conclusion

We described a human study to assess video quality which was conducted on multiple mobile platforms and encompassed a wide variety of dis-

	Comp.	RA	TD	WL	All
PSNR	0.8331	0.1365	0.0342	0.8391	0.3821
SS-SSIM	0.7570	0.0212	0.0302	0.7722	0.3526
MS-SSIM	0.7959	0.2384	0.0327	0.8589	0.4010
VSNR	0.9764	0.2054	0.0360	1.0432	0.4614
VIF	1.0555	0.2094	0.0022	**1.1661**	0.4959
UQI	0.4549	0.0407	0.0128	0.7934	0.3507
NQM	0.7845	0.2172	0.0262	1.1043	0.4651
WSNR	0.7739	0.1365	0.0004	0.7658	0.3197
SNR	0.5727	0.0755	0.0014	0.5297	0.2156
VQM	0.7966	0.2337	0.0370	0.8392	0.3830
MOVIE	0.8897	0.2201	0.0117	0.7635	0.4100
Threshold F-ratio	1.1390	1.1622	1.1234	1.1390	1.0672

Table 5.18: Mobile Study: Algorithm performance vs. the theoretical null model. Listed are the F-ratios i.e., ratio of (a) variances of residuals between the differential opinion scores (DOS) and algorithm scores and (b) variances of residuals between the differential opinion scores (DOS) and DMOS for each distortion as well as across all distortions. Also listed is the threshold F-ratio. The algorithm is statistically equivalent to the null model if the F-ratio is greater than the threshold F-ratio. Bold font indicates statistical equivalence to the theoretical null model.

	Comp.	RA	TD	WL	All
PSNR	0.9773	0.0859	0.0043	0.6947	0.2932
SS-SSIM	0.5638	0.0802	0.0005	0.4514	0.1743
MS-SSIM	0.8095	0.1434	0.0031	0.6463	0.2809
VSNR	0.9873	0.1163	0.0033	0.6930	0.3022
VIF	1.1904	0.1589	0.0002	0.9459	0.4242
UQI	0.2844	0.0271	0.0063	0.4224	0.0771
NQM	1.0823	0.0766	0.0009	0.8928	0.3749
WSNR	1.0915	0.2023	0.0032	0.6820	0.3233
SNR	0.8421	0.1623	0.0084	0.4884	0.2237
VQM	0.7773	0.0717	0.0000	0.6280	0.2462
MOVIE	1.1253	0.2897	0.0000	0.9966	0.4260
Threshold F-ratio	1.1956	1.2292	1.1732	1.1956	1.0831

Table 5.19: Tablet Study: Algorithm performance vs. the theoretical null model. Listed are the F-ratios i.e., ratio of (a) variances of residuals between the differential opinion scores (DOS) and algorithm scores and (b) variances of residuals between the differential opinion scores (DOS) and DMOS for each distortion as well as across all distortions. Also listed is the threshold F-ratio. The algorithm is statistically equivalent to the null model if the F-ratio is greater than the threshold F-ratio. Bold font indicates statistical equivalence to the theoretical null model.

tortions, including dynamically-varying distortions as well as uniform compression and wireless packet-loss. The large size of the study and the variety that it offers allows one to study and analyze human reactions to temporally varying distortions as well as to varying form factors from a wide variety of perspectives. We make a number of further observations that may prove – from the perspective of understanding human reactions to complex, time varying distortions and from the algorithm design perspective.

An obvious conclusion from our analysis is that time-varying quality has a definite impact on human subjective judgments of quality, and this impact is a function of the frequency of occurrence of significant distortion changes and of the differences in quality between segments. Humans seemingly prefer longer freezes over shorter ones – this is not terribly surprising since choppy video playback is not pleasing at all. However, what is surprising about the frame-freeze distortion is that humans appear to be far more forgiving of lost segments than they are of choppy quality. This has interesting implications for those supplying real-time video delivery. It is also prudent to note that while choppy playback is the worst offender, lost segments start to matter relative to small reductions in choppiness. Further, this preference is dependent upon the content being displayed. It would be interesting to study whether the same results hold true when viewing sports – a viewer may prefer choppy playback in this case as opposed to him missing out on the footage of that all important goal being scored. On the flip side, in applications such as video chatting it is possible that our results will be further validated. The data in this study seems

to indicate that designers should use algorithms for resource allocation that penalize semi-filled buffers over those that penalize completely empty buffers.

The data from the rate adaptation and temporal dynamics distortions, while somewhat contrary to popularly held notions on human perception of quality are intuitive and interesting. The first observation is that humans are not as unforgiving as one would imagine them to be. In fact they seem to reward attempts to improve quality. As we summarized in the temporal dynamics discussion, when the user is subjected to a long spell of good quality video, s/he has seemingly taken that level of quality for granted, and when the provider switches to a much lower quality level, he is severe with his rating of quality. On the contrary, faster rate changes seemingly push the user to believe that the provider is attempting to maximize his quality of experience and hence these videos are given higher quality scores. Another explanation is that less rapid rate changes can produce long periods of low-quality video bracketed by segments of high-quality videos. In this case, the low quality may be regarded as more enduring, and hence, more annoying. Due to the limitations of study sessions we were unable to include the other condition – $R_4 - R_1 - R4$ – here, not only is there high quality at the end, but there is also a segment of poor quality in the middle. From the current data it is difficult to predict how the user may react to this situation. Of course, variations on the rate of fluctuation in quality is another area to explore.

The field of analyzing continuous-time human opinion scores of quality is one that is still nascent. We explored a small set of preliminary temporal

pooling ideas drawn from the literature or from conventional wisdom. Our results, while encouraging, still do not completely explain human responses to temporally varying distortions. For compression and wireless distortions, the mean of human opinion across time is a good indicator of the final quality – possibly owing to the fact that with stagnant quality, the human simply picks the mean when providing continuous quality scores. What is surprising is the performance of percentile pooling. This strategy works well for larger screen displays (albeit using an indirect method to assess its performance – pooling of objective scores [175, 219, 315]), but humans are seemingly more forgiving of poorer quality when viewing videos on smaller form factors. The observations from the memory-effect pooling are intriguing. While the mean of continuous quality scores is poor indicator of the final quality for videos with dynamically varying distortions, memory-effect based pooling seems to better capture human responses. With a change in the device form factor however, even this pooling strategy begins to fail. This implies that there is a lot more work to be done in understanding how humans integrate continuous quality scores and produce the final summarized score that they give each video. This is even more true for the frame-freeze distortions. It is unclear at this point how humans rate the effects of frame-freeze distortions on the temporal perception of video quality.

While a lot more can be said with regards to the human data, in the interest of space we now move our discussion to the objective algorithms. To us, the main takeaway from the analysis is that scalability, which requires

multiscale processing, is a desirable property to assess the quality of videos of diverse sizes, resolutions and display forms. Single-scale algorithms such as VQM and SS-SSIM, which do well on videos shown on larger screens, may not accurately predict the quality of videos displayed on smaller screens.

Results from the temporally varying distortions are both disappointing and encouraging at the same time. It seems that for smaller rate variations, the algorithms manage to do resonably well in predicting quality, however with increased variation in the temporal distortion patterns, the algorithms fail. While this may be due to a multitude of factors, one possible reason could be the temporal pooling strategy applied. For the IQA algorithms, our strategy was simply to use the temporal mean of the frame-level scores, while the VQA algorithms pooled the predicted temporal scores as they were designed to do (eg., MOVIE uses the mean). In light of the results from our temporal pooling analysis of human scores and recent research in temporal pooling strategies for objective algorithms [210, 253], it seems very likely that algorithm performance can be improved by employing more appropriate strategies for integrating quality scores over time. Incorporating knowledge of the device and human responses to temporal quality as a function of the form factor should lead to additional benefits. Clearly, there remains ample room for developing better VQA algorithms – since none of the algorithms are equivalent (or even close) to the theoretical null model.

We hope that the new LIVE mobile VQA database of 200 distorted

videos and associated human opinion scores from over 50 subjects will provide fertile ground for years of future research. Given the sheer quantity of data, we believe that our foregoing analysis is the tip of the ice-berg of discovery. We invite further analysis of the data towards understanding and producing better models of human behavior when viewing videos on mobile platforms. Other fields of inquiry that may benefit from this database include human behavior modeling; application and content driven analysis of human behavior; device and context-specific design of objective algorithms; video network resource allocation over time and many others. Given the explosion of mobile devices, and associated load on bandwidth, we believe that the work presented here and the observations made with regards to human behavior will serve as essential tools in modeling video delivery over wireless networks.

Chapter 6

Conclusion and Future Work

In this dissertation we detailed a no-reference (NR) image quality assessment (IQA) algorithm based on natural scene statistics (NSS) that first identifies the distortion present in the image and then proceeds to perform blind quality assessment. We demonstrated a novel application of such a two-stage framework – blind, perceptually optimized, general purpose image repair. Having described the NR QA problem and its application, we also showed that the distortion-agnostic NR algorithm is statistically indistinguishable from a leading full-reference (FR) IQA algorithm, and that image repair using this framework produces noticeable improvements in quality. In the final section of this dissertation, we described a large-scale human study that we conducted to assess human behavior and opinion on quality of videos viewed on small-resolution screens such as mobile phones and tablets. The distortions simulated included the previously studied uniform compression and wireless packet loss and the novel dynamically-varying distortions. The large size of the study and the variety that it offers allows one to study and analyze human reactions to temporally varying distortions as well as to varying form factors from a wide variety of perspectives. We have made some comments on the future of each of these areas, and below we summarize some more interesting avenues

for future work. The interested reader is directed to [176] for a more detailed exposition.

No-reference video quality assessment (NR VQA) is a difficult problem to solve. Even though a host of methods have been proposed in literature, most of these methods are only for the full-reference case. As of this writing, there does not exist an algorithm for NR VQA that is capable of assessing more than one distortion. Most researchers tend to select a particular kind of distortion that affects videos and evaluate quality. Any naive viewer of videos will testify to the fact that distortions in videos are not singular. In fact, compression – which is generally assumed to have a blocking distortion, also introduces blurring and motion- compensation mismatches, mosquito noise, ringing and so on [338]. Given that there exist a host of distortions that may affect a video, one should question the virtue of trying to model each individual distortion. Further, if one does choose to model each distortion individually, a method to study the effect of multiple distortions must be undertaken. Again, this is a combinatorially challenging problem.

A majority of algorithms seek to model spatial distortions alone and even though some methods include elementary temporal features, a wholesome approach to NR VQA should involve a spatio-temporal distortion model. Further, in most cases a majority of the design decisions are far removed from human vision processing. It is imperative as researchers that we keep in mind that the ultimate receiver is the human and hence understanding and incorporating HVS properties in an algorithm is of essence. It is our belief that

NSS-based approaches, such as those discussed here for NR IQA, will allow for the development of successful NR VQA algorithms and the development of NR VQA algorithms remains an exciting research problem for the future.

Our general outlook towards quality assessment is one of cautious optimism – we believe that a lot more needs to be done in the area of quality assessment and that each of the subfields described here (as well as those that were not) have great potential for growth. Our stance is that algorithmic quality assessment will heavily benefit from research in visual psychophysics. As we understand the human visual system better, quality assessment algorithms that incorporate these mechanisms will surely result in better performance.

Apart from quality assessment, there exist related esoteric topics such as aesthetics assessment, 3D perception, and the effect that scene content can have on user opinion. Multi-modal (multimedia) quality assessment remains exciting as well.

A famous African proverb goes 'Tomorrow belongs to the people who prepare for it today' and we hope that we have at least started preparing for this journey and that this dissertation has laid a solid foundation for the future of quality assessment research.

Appendices

Appendix A

Mapping MS-SSIM to DMOS

Instead of directly using the MS-SSIM scores to quantify quality, we re-map the MS-SSIM scores to the more easily interpreted perceptual scale of differential mean opinion scores (DMOS), obtained from human subjective studies such as that in [264]. We use the human DMOS obtained from [264], and map MS-SSIM scores via a logistic function fit (A.1), where the parameters $\beta_i \quad \{i = 1, 2, \ldots, 5\}$ are estimated via a nonlinear optimization procedure (MATLAB function `nlinfit`) between the DMOS and the MS-SSIM scores. This was done for each of the four distortions targeted by GENII-1, to produce MS-SSIM$_D$.

$$f(x) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + exp\left(-\beta_2 (x - \beta_3)\right)} \right] + \beta_4 x + \beta_5 \tag{A.1}$$

The non-linear fitting procedure detailed here is identical to that used in [264] prior to computating linear correlation and root-mean squared error between algorithm scores and DMOS.

While such a remap using a database is limited by the database and is specific to it, the LIVE IQA database of [264] incorporates a wide variety of distortions levels and spans a good range of visual quality and hence, the

re-mapped scores obtained are reasonable representations of visual quality on the linear DMOS scale, where 0 indicates the best quality and 100 indicates the worst quality.

Appendix B

Instructions to the Subject

You are taking part in a study to assess the quality of videos. You will be shown a video at the center of your screen and there will be a rating bar at the bottom, which can be controlled by using your fingers on the touchscreen. You are to provide the quality as function of time – i.e., move the rating bar in real-time based on your instantaneous perception of quality. The extreme left on the bar is bad quality and the extreme right is excellent quality. At the end of the video you will be presented with a similar bar, this time calibrated as 'Bad', 'Poor' and 'Excellent', from left-to-right. Using this bar, provide us with your opinion on the overall quality of the video. There is no right or wrong answer, we simply wish to gauge your opinion on the quality of the video that is shown to you.

Bibliography

[1] Common Test Conditions for RTP/IP over 3GPP/3GPP2. *http://ftp3.itu.ch/av-arch/videosite/0109 San/VCEG-N80_software.zip*.

[2] Toyama image database. *http://mict.eng.u-toyama.ac.jp/mict/index2.html*.

[3] Video quality metric. *http://www.its.bldrdoc.gov/n3/video/VQM_software.php*.

[4] Visual signal to noise ratio. *http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr_matlab_source.zip*.

[5] Advanced video coding. *ISO/IEC 14496-10 and ITU-T Rec. H.264*, 2003.

[6] H.264/MPEG-4 AVC reference software manual. *http://iphome.hhi.de/suehring/tml/JM_Reference_Software_Manual_(JVT-X072).pdf*, 2007.

[7] The XGL Toolbox. *http://128.83.207.86/jsp/software/xgltoolbox-1.0.5.zip*, 2008.

[8] Live Video Database. *http://live.ece.utexas.edu/research/quality/live_video.html*, 2009.

[9] A. J. Ahumada Jr. Computational image quality metrics: A review. *SID Digest*, 24:305–308, 1993.

[10] A. Albonico, G. Valenzise, M. Naccari, M. Tagliasacchi, and S. Tubaro. A Reduced-Reference Video Structural Similarity metric based on no-reference estimation of channel-induced distortion. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009.

[11] H. C. Andrews and B. R. Hunt. *Digital image restoration*, volume 1. Prentice-Hall Signal Processing Series, Englewood Cliffs: Prentice-Hall, 1977.

[12] World Airline Entertainment Association. Digital content delivery methodology for airline in-flight entertainment systems.

[13] J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2):213–251, 1992.

[14] R.V. Babu, A.S. Bopardikar, A. Perkis, and O.I. Hillestad. No-reference metrics for video streaming applications. *International Packet Video Workshop*, 2004.

[15] M.R. Banham and A.K. Katsaggelos. Digital image restoration. *IEEE Signal Processing Magazine*, 14(2):24–41, 1997.

[16] J. Bardsley, S. Jefferies, J. Nagy, and R. Plemmons. Blind iterative restoration of images with spatially-varying blur. *Optics Express*, 14(1767-1782):2, 2006.

[17] M. Barkowsky, B. Eskofier J. Bialkowski and, R. Bitto, and A. Kaup. Temporal trajectory aware video quality measure. *IEEE Jnl. Sp. Top. Sig. Proc.*, 3(2):266–279, April 2009.

[18] R. Barland and A. Saadane. Reference free quality metric for JPEG-2000 compressed images. In *Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on*, volume 1, 2005.

[19] R. Barland and A. Saadane. A reference free quality metric for compressed images. *Proc. of 2nd Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2006.

[20] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Computing Surveys (CSUR)*, 27(3):433–466, 1995.

[21] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63:75–104, 1996.

[22] R.T. Born and D.C. Bradley. Structure and Function of Visual Area MT. *Annual Review of Neuroscience*, 28:157–189, 2005.

[23] P. Bourdon, B. Augereau, C. Olivier, and C. Chatellier. A PDE-based method for ringing artifact removal on grayscale and color JPEG2000 images. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[24] A. C. Bovik and Z. Wang. *Modern Image Quality Assessment*. Morgan and Claypool Publishing Co, New York, 2006.

[25] A.C. Bovik. Meditations on Visual Quality. *IEEE COMPSOC e-letter, Technology Advances*, May 2009.

[26] T. Brox, O. Kleinschmidt, and D. Cremers. Efficient nonlocal means for denoising of textural patterns. *Image Processing, IEEE Transactions on*, 17(7):1083–1092, 2008.

[27] O. Bryt and M. Elad. Improving the k-SVD Facial Image Compression using a Linear Deblocking Method. In *IEEE Convention of Electrical and Electronics Engineers in Israel*, pages 533–537, 2008.

[28] BT. 500-11:Methodology for the subjective assessment of the quality of television pictures.. *International Telecommuncation Union, Geneva, Switzerland*, 2002.

[29] A. Buades, B. Coll, and J. M. Morel. A Review of Image Denoising Algorithms, with a New One. *Multiscale Modeling and Simulation*, 4(2):490–530, 2006.

[30] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[31] J. F. Cai, S. Osher, and Z. Shen. Linearized bregman iterations for frame-based image deblurring. *SIAM Journal on Imaging Sciences*, 2:226–252, 2009.

[32] P. Campisi, M. Carli, G. Giunta, and A. Neri. Blind quality assessment system for multimedia communications using tracing watermarking. *IEEE Transactions on Signal Processing*, 51(4):996–1002, 2003.

[33] M. Carandini, J.B. Demb, V. Mante, D.J. Tolhurst, Y. Dan, B.A. Olshausen, J.L. Gallant, and N.C. Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, 2005.

[34] M. Carli, M. C. Q. Farias, E. D. Gelasca, R. Tedesco, and A. Neri. Quality assessment using data hiding on perceptually important areas. *IEEE International Conference on Image Processing*, 2005.

[35] A. Cavallaro and S. Winkler. Segmentation-driven perceptual quality metrics. In *Image Processing, 2004. ICIP'04. 2004 International Conference on*, volume 5, 2004.

[36] J. Caviedes and S. Gurbuz. No-reference sharpness metric based on local edge kurtosis. In *Proc. of IEEE Int. Conf. on Image Processing*, volume 3, pages 53–56, 2002.

[37] J. Caviedes and J. Jung. No-reference metric for a video quality control loop. *5th World Multiconference on Systemics, Cybernetics and Informatics*, 2001.

[38] J. Caviedes and F. Oberti. A new sharpness metric based on local kurtosis, edge and energy information. *Signal Processing: Image Communication*, 19(2):147–161, 2004.

[39] J.E. Caviedes and F. Oberti. No-reference quality metric for degraded and enhanced video. *Proceedings of SPIE*, 5150:621, 2003.

[40] D. M. Chandler and S. S. Hemami. A57 database. *http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html*, 2007.

[41] D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, 2007.

[42] C.C. Chang and C.J. Lin. LIBSVM: a library for support vector machines. *http://www.csie.ntu.edu.tw/ cjlin/libsvm/*, 2001.

[43] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath. Design of linear equalizers optimized for the structural similarity index. *IEEE Transactions on Image Processing*, 17(6):857–872, 2008.

[44] S. S. Channappayya, A. C. Bovik, and R. W. Heath. A linear estimator optimized for the structural similarity index and its application to image denoising. In *IEEE International Conference on Image Processing*, pages 2637–2640. IEEE, 2006.

[45] S. S. Channappayya, A. C. Bovik, and R. W. Heath. Rate bounds on SSIM index of quantized images. *IEEE Transactions on Image Processing*, 17(9):1624–1639, 2008.

[46] C. Charrier, K. Knoblauch, A. K. Moorthy, A. C. Bovik, and L. T. Maloney. Comparison of image quality assessment algorithms on com-

pressed images. *SPIE conference on Image quality and System Performance*, 2010.

[47] C. Charrier, L. T. Maloney, H. Cheri, and K. Knoblauch. Maximum likelihood difference scaling of image quality in compression-degraded images. *Journal of the Optical Society of America*, 24(11):3418 – 3426, 2007.

[48] P. Chatterjee and P. Milanfar. Clustering-based denoising with locally learned dictionaries. *Transactions on Image Processing*, 18(7):1438–1451, 2009.

[49] P. Chatterjee and P. Milanfar. Is denoising dead? *IEEE Transactions on Image Processing*, 19(4):895–911, 2010.

[50] G.H. Chen, C.L. Yang, and S.L. Xie. Gradient-based structural similarity for image quality assessment. *IEEE International Conference on Image Processing*, pages 2929–2932, 2006.

[51] J. Chen, T.N. Pappas, A. Mojsilovic, and B. Rogowitz. Adaptive image segmentation based on color and texture. *IEEE Intl. Conf. Image Proc.*, 2:789–792, 2002.

[52] J. Chen, Y. Zhang, L. Liang, S. Ma, R. Wang, and W. Gao. A No-Reference Blocking Artifacts Metric Using Selective Gradient and Plainness Measures. In *Proc. of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, 2008.

[53] M. J. Chen and A. C. Bovik. No-reference Image Blur Assessment using Multiscale Gradient. *1st International Workshop on Quality of Multimedia Experience (QoMEX)*, 2009.

[54] M. J. Chen and A. C. Bovik. Fast structural similarity index algorithm. *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 994–997, 2010.

[55] Y. Y. Chen, Y. W. Chang, and W. C. Yen. Design of a de-ringing filter for wavelet-based compressed image. In *International Technical Conference on Circuits/Systems, Computers and Communications*, pages 1265–1268, 2008.

[56] Y.Y. Chen, S.C. Tai, C.X. Wang, and K.W. Lin. Design of a filter against artifacts for JPEG2000. *Journal of Electronic Imaging*, 14:043002, 2005.

[57] CISCO Corp. Cisco visual networking index: Global mobile data traffic forecast update, 2010–2015. *http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html*, 2011.

[58] C. Cotsaces, N. Nikolaidis, and I. Pitas. Video shot detection and condensed representation. A Review. *IEEE signal processing magazine*, 23(2):28–37, 2006.

[59] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot. An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. *Medical Imaging, IEEE Transactions on*, 27(4):425–441, 2008.

[60] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[61] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

[62] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image restoration by sparse 3D transform-domain collaborative filtering. In *SPIE Electronic Imaging*. Citeseer, 2008.

[63] S. Daly. Engineering observations from spatiovelocity and spatiotemporal visual models. *Proc. of SPIE*, 3299:180–191, 1998.

[64] S.J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. *Proceedings of SPIE*, 1666:2, 1992.

[65] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, 2002.

[66] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial Mathematics, 1992.

[67] J.G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, 1985.

[68] A. Dauwe, B. Goossens, H. Q. Luong, and W. Philips. A fast non-local image denoising algorithm. In *Proceedings of SPIE*, volume 6812, page 681210, 2008.

[69] C.A. Deledalle, F. Tupin, and L. Denis. Poisson nl means: Unsupervised non local means for poisson noise. In *International Conference on Image Processing*, pages 801–804. IEEE, 2010.

[70] D. Dong and J. Atick. Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6(2):159–178, 1995.

[71] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[72] R. Dosselmann and X.D. Yang. A Prototype No-Reference Video Quality System. In *Computer and Robot Vision, 2007. CRV'07. Fourth Canadian Conference on*, pages 411–417, 2007.

[73] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley Interscience, New York, 2001.

[74] A. Eichhorn and P. Ni. Pick your layers wisely-a quality assessment of h. 264 scalable video coding for mobile devices. *Proceedings of the 2009 IEEE international conference on Communications*, pages 5446–5451, 2009.

[75] M. Elad. On the origin of the bilateral filter and ways to improve it. *IEEE Transactions on Image Processing*, 11(10):1141–1151, 2002.

[76] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[77] E. ETSI. 302 304 V1. 1.1, Digital Video Broadcasting (DVB): Transmission System for Handheld Terminals (DVB-H), December 2004.

[78] M.C.Q. Farias, M. Carli, A. Neri, and S.K. Mitra. Video quality assessment based on data hiding driven by optical flow information. *Proceedings of the SPIE Human Vision and Electronic Imaging IX, San Jose, CA, USA*, pages 190–200, 2004.

[79] M.C.Q. Farias and S.K. Mitra. No-reference video quality metric based on artifact measurements. In *IEEE International Conference on Image Processing*, volume 3, pages 141–144, 2005.

[80] M.C.Q. Farias, M.S. Moore, J.M. Foley, and S.K. Mitra. Perceptual contributions of blocky, blurry, and fuzzy impairments to overall an-

noyance. *Proceedings of the IS&T/SPIE Human Vision and Electronic Imaging IX*, 5292:109–120, 2004.

[81] R. Ferzli and L.J. Karam. A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB). *IEEE Transactions on Image Processing*, 18(4):717, 2009.

[82] D.J. Fleet and A.D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, 1990.

[83] A. Foi, V. Katkovnik, and K. Egiazarian. Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. *Image Processing, IEEE Transactions on*, 16(5):1395–1411, 2007.

[84] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comp. Sys. Sc.*, 55(1):119–139, 1997.

[85] SM Friend and CL Baker. Spatio-temporal frequency separability in area 18 neurons of the cat. *Vision research*, 33:1765–1771, 1993.

[86] Y. Fu-zheng, W. Xin-dai, C. Yi-lin, and W. Shuai. A no-reference video quality assessment method based on digital watermark. *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003*, 3, 2003.

[87] B. Furht and S.A. Ahson. *Handbook of Mobile Broadcasting: Dvb-h, Dmb, Isdb-t, and Mediaflo.* Auerbach Publications, 2008.

[88] S. Gabarda and G. Cristobal. Blind image quality assessment through anisotropy. *J. Opt. Soc. Am. A*, 24:B42–B51, 2007.

[89] S. Gabarda and G. Cristobal. *http://www.iv.optica.csic.es/page49/page51/page51.html*, November 2010.

[90] X. Gao, T. Wang, and J. Li. A content-based image quality metric. *Lecture Notes in Computer Science*, 3642:231, 2005.

[91] P. Gastaldo, R. Zunino, I. Heynderickx, and E. Vicario. Objective quality assessment of displayed images by using neural networks. *Signal Processing: Image Communication*, 20(7):643–661, 2005.

[92] Matthew Gaubatz. Metrix mux visual quality assessment package. *http://foulard.ece.cornell.edu/gaubatz/metrix_mux/*.

[93] W.S. Geisler. Visual perception and the statistical properties of natural scenes. *Ann. Rev. Neuroscience*, 2007.

[94] B. Girod. What's wrong with mean-squared error?, Digital images and human vision, A. B. Watson, Ed. pages 207–220, 1993.

[95] R.C. Gonzalez and R.E. Woods. *Digital image processing.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA,, 2002.

[96] B. Goossens, Q. Luong, A. Pizurica, and W. Philips. An improved non-local denoising algorithm. In *Local and Non-Local Approximation in Image Processing, International Workshop, Proceedings*, page 143, 2008.

[97] J. A. Guerrero-Colón and J. Portilla. Deblurring-by-denoising using spatially adaptive gaussian scale mixtures in overcomplete pyramids. In *IEEE International Conference on Image Processing*, pages 625–628, 2006.

[98] J.A. Guerrero-Colón, L. Mancera, and J. Portilla. Image restoration using space-variant gaussian scale mixtures in overcomplete pyramids. *Image Processing, IEEE Transactions on*, 17(1):27–41, 2008.

[99] S. R. Gulliver and G. Ghinea. The perceptual and attentive impact of delay and jitter in multimedia delivery. *Broadcasting, IEEE Transactions on*, 53(2):449–458, 2007.

[100] I. P. Gunawan and M. Ghanbari. Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(1):71–83, 2008.

[101] S. Gupta, M. K. Markey, and A. C. Bovik. Advances and challenges in 3D and 2D+ 3D human face recognition. In *Pattern recognition in biology*. Nova Science Publishers, 2007.

[102] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[103] J.F. Hair. *Multivariate data analysis*. Prentice Hall, 2006.

[104] D. Hands, A. Bourret, and D. Bayart. Video QoS enhancement using perceptual quality metrics. *BT Technology Journal*, 23(2):208–216, 2005.

[105] S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall, 2008.

[106] A. P. Hekstra, J. G. Beerends, D. Ledermann, F. E. De Caluwe, S. Kohler, R. H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss. PVQM–a perceptual video quality measure. *Signal Processing: Image Communication*, 17(10):781–798, 2002.

[107] S. Higginbotham. Spectrum shortage will strike in 2013. *http://gigaom.com/2010/02/17/analyst-spectrum-shortage-will-strike-in-2013/*, 2010.

[108] B. Hiremath, Q. Li, and Z. Wang. Quality-aware video. *IEEE International Conference on Image Processing*, 3, 2007.

[109] C.W. Hsu, C.C. Chang, C.J. Lin, et al. A practical guide to support vector classification. *http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf*, 2003.

[110] Y. H. Huang, T. S. Ou, P. Y. Su, and H. H. Chen. Perceptual rate-distortion optimization using structural similarity index as quality metric. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(11):1614–1624, 2010.

[111] D.H. Hubel, J. Wensveen, and B. Wick. Eye, brain, and vision. 1988.

[112] Q. Huynh-Thu and M. Ghanbari. Impact of jitter and jerkiness on perceived video quality. In *Proceedings of the Workshop on Video Processing and Quality Metrics*. Citeseer, 2006.

[113] International Standards Organization (ISO). *http://www.iso.org/iso/home.htm*.

[114] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 2002.

[115] ITU-T and ISO/IEC JTC 1. ITU-T Recommendation H.262 and ISO/IEC 13 818-2 (MPEG-2). Generic Coding of Moving Pictures and Associated Audio Information - Part 2: Video. 1994.

[116] Joint Video Team (JVT). SVC Reference Software (JSVM software). *http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm*.

[117] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG. Joint Draft ITU-T Rec. H.264 — ISO/IEC 14496-10 / Amd.3 Scalable video

coding. *http://www.hhi.fraunhofer.de/fileadmin/hhi/downloads/IP/ip_ic_H.264-MPEG4-AVC-Version8-FinalDraft.pdf*.

[118] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG. Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14 496-10 AVC. *JVT-G050*, 2003.

[119] N. Joshi, C. L. Zitnick, R. Szeliski, and D. J. Kriegman. Image deblurring and denoising using color priors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1550–1557, 2009.

[120] S. Jumisko-Pyykko and J. Hakkinen. Evaluation of subjective video quality of mobile devices. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 535–538. ACM, 2005.

[121] S. Jumisko-Pyykko and M.M. Hannuksela. Does context matter in quality evaluation of mobile television? In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 63–72. ACM, 2008.

[122] JVT. H.264/AVC software coordination. *http://iphome.hhi.de/suehring/tml/*.

[123] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan. Modeling packet-loss visibility in MPEG-2 video. *IEEE transactions on Multimedia*, 8(2):341–355, 2006.

[124] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for image artifacts based on human visualsensitivity. *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94.*, 1994.

[125] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for blocking artifacts in images based on human visual sensitivity. *IEEE Transactions on image processing*, 4(6):713–724, 1995.

[126] Y. Kawayoke and Y. Horita. NR objective continuous video quality assessment model based on frame quality measure. In *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, pages 385–388, 2008.

[127] V. Kayargadde and J.B. Martens. Perceptual characterization of images degraded by blur and noise: model. *Journal of the Optical Society of America A*, 13:1178–1188, 1996.

[128] C. Keimel, T. Oelbaum, and K. Diepold. No-Reference Video Quality Evaluation for High-Definition Video. In *Proceedings of the International Conference on Image Processing, San Diego, CA, USA*, 2009.

[129] D. H. Kelly. Spatiotemporal variation of chromatic and achromatic contrast thresholds. *Journal of the Optical Society of America*, 73(6):742–750, 1983.

[130] C. Kervrann and J. Boulanger. Optimal spatial adaptation for patch-based image denoising. *Transactions on Image Processing*, 15(10):2866–2878, 2006.

[131] V. Khryashchev, I. Apalkov, and L. Shmaglit. A novel smart bilateral filter for ringing artifacts removal in JPEG2000 images. In *International Conference on Image Processing, Computer Vision and Pattern Recognition*, 2010.

[132] M. Knee. A single-ended picture quality measure for MPEG-2. *Proc. International Broadcasting Convention*, pages 7–12, 2000.

[133] H. Knoche, J. D. McCarthy, and M. A. Sasse. Can small be beautiful?: assessing image resolution requirements for mobile tv. *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 829–838, 2005.

[134] A. Krylov and A. Nasonov. Adaptive total variation deringing method for image interpolation. In *IEEE International Conference on Image Processing*, pages 2608–2611, 2008.

[135] P. Le Callet, C. Viard-Gaudin, and D. Barba. A convolutional neural network approach for objective video quality assessment. *IEEE Transactions on Neural Networks*, 17(5):1316, 2006.

[136] Patrick Le Callet and Florent Autrusseau. Subjective quality assessment irccyn/ivc database, 2005. http://www.irccyn.ec-nantes.fr/ivcdb/.

[137] E. Le Pennec and S. Mallat. Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438, 2005.

[138] A. Leontaris, P. C. Cosman, and A. R. Reibman. Quality evaluation of motion-compensated edge artifacts in compressed video. *IEEE Transactions on Image Processing*, 16(4):943–956, 2007.

[139] A. Leontaris and A. R. Reibman. Comparison of blocking and blurring metrics for video compression. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, volume 2, 2005.

[140] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Deconvolution using natural image priors. *ACM SIGGRAPH*, 2007.

[141] C. Li and A.C. Bovik. Three-Component Weighted Structural Similarity Index. *Proceedings of SPIE*, 7242:72420Q, 2009.

[142] Q. Li and Z. Wang. Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation. *IEEE J. Selected Topics in Signal Proc.*, 3(2):202–211, 2009.

[143] X. Li. Blind image quality assessment. In *Intl. Conf. on Image Processing, New York, USA*, 2002.

[144] X. Li. Improved wavelet decoding via set theoretic estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):108–112, 2005.

[145] X. Li. Fine-granularity and spatially-adaptive regularization for projection-based image deblurring. *Image Processing, IEEE Transactions on*, (99):1–1, 2011.

[146] Z. Li and E. J. Delp. Block artifact reduction using a transform-domain markov random field model. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(12):1583–1593, 2005.

[147] A. W. C. Liew, H. Yan, and N. F. Law. POCS-based blocking artifacts suppression using a smoothness constraint set with explicit region modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(6):795–800, 2005.

[148] A.C. Likas and N.P. Galatsanos. A variational approach for Bayesian blind image deconvolution. *IEEE Tran. Signal Proc.*, 52(8):2222–2233, 2004.

[149] H. Liu and I. Heynderickx. A perceptually relevant no-reference blockiness metric based on local image characteristics. *EURASIP Journal on Advances in Signal Processing*, 2009, 2009.

[150] T. Liu, Y. Wang, J. M. Boyce, H. Yang, and Z. Wu. A novel video quality metric for low bit-rate video considering both coding and packet-

loss artifacts. *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3(2), April 2009.

[151] J. Lu. Image analysis for video artifact estimation and measurement. *Proceedings of SPIE*, 4301:166, 2001.

[152] Z. Lu, W. Lin, B.C. Seng, S. Kato, E. Ong, and S. Yao. Perceptual Quality Evaluation on Periodic Frame-Dropping Video. *Proc. of IEEE Conference on Image Processing*, pages 433–436, 2007.

[153] J. Lubin. A visual discrimination model for imaging system design and evaluation. *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*, page 245, 1995.

[154] J. Lubin and D. Fibush. Sarnoff JND vision model. *T1A1*, 5:97–612, 1997.

[155] F. J. MacWilliams and N. J. A. Sloane. Pseudo-random sequences and arrays. *Proceedings of the IEEE*, 64(12):1715–1729, 1976.

[156] M. Mahmoudi and G. Sapiro. Fast image and video denoising via non-local means of similar neighborhoods. *Signal Processing Letters, IEEE*, 12(12):839–842, 2005.

[157] S. G. Mallat. *A wavelet tour of signal processing*. Academic Pr, 1999.

[158] L.T. Maloney and J.N. Yang. Maximum likelihood difference scaling. *Journal of Vision*, 3(8):573–585, 2003.

[159] J. Mannos and D. Sakrison. The effects of a visual fidelity criterion on the encoding of images. *IEEE Trans. Inf. Theory*, 20(4):525–535, 1974.

[160] V. Mante, V. Bonin, and M. Carandini. Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron*, 58:625–638, May 2008.

[161] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[162] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. A no-reference perceptual blur metric. *IEEE Int'l Conf. Image Proc.*, 3:57–60, 2002.

[163] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. Perceptual blur and ringing metrics: Application to JPEG2000. *Signal Processing: Image Communication*, 19(2):163–172, 2004.

[164] M. Masry, S. S. Hemami, and Y. Sermadevi. A scalable wavelet-based video distortion metric and applications. *IEEE Tran. Cir. Sys. Vid. Tech.*, 16(2):260–273, 2006.

[165] F. Massidda, D.D. Giusto, and C. Perra. No reference video quality estimation based on human visual system for 2.5/3G devices. In *Proceedings of SPIE*, volume 5666, page 168, 2005.

[166] P. Meer. Robust techniques for computer vision. *Emerging topics in computer vision*, pages 107–190, 2004.

[167] L. Meesters and J. B. Martens. A single-ended blockiness measure for JPEG-coded images. *Signal Processing*, 82(3):369–387, 2002.

[168] F. Meng, X. Jiang, H. Sun, and S. Yang. Objective Perceptual Video Quality Measurement using a Foveation-Based Reduced Reference Algorithm. *IEEE International Conference on Multimedia and Expo*, pages 308–311, 2007.

[169] A. Mittal, A. K. Moorthy, and A. C. Bovik. Blind/referenceless image spatial quality evaluator. In *Asilomar Conference on Signals, Systems and Computers*, November 2011.

[170] A. Mittal, A. K. Moorthy, and A. C. Bovik. General-purpose blind image quality assessment in the spatial domain. *IEEE Transactions Image Processing*, 2011 (submitted).

[171] A. Mittal, A. K. Moorthy, and A. C. Bovik. Automatic parameter prediction for image denoising algorithms using perceptual quality features. In *SPIE Proceedings Human Vision and Electronic Imaging*, 2012.

[172] D. C. Montgomery and G. C. Runger. *Applied Statistics and Probability for Engineers*. Wiley-Interscience, 1999.

[173] A. K. Moorthy and A. C. Bovik. A motion-compensated approach to video quality assessment. *Proc. IEEE Asilomar Conference on Signals, Systems and Computers*, 2009.

[174] A. K. Moorthy and A. C. Bovik. Perceptually significant spatial pooling techniques for image quality assessment. *Human Vision and Electronic Imaging XIV. Proceedings of the SPIE*, 7240, January 2009.

[175] A. K. Moorthy and A. C. Bovik. Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3(2):193–201, April 2009.

[176] A. K. Moorthy and A. C. Bovik. Automatic prediction of perceptual video quality: Recent trends and research directions. *High-Quality Visual Experience*, pages 3–23, 2010.

[177] A. K. Moorthy and A. C. Bovik. Statistics of natural image distortions. *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 962–965, 2010.

[178] A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(2):587–599, May 2010.

[179] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From

scene statistics to perceptual quality. *IEEE Transactions Image Processing*, 20(12):3350–3364, December 2011.

[180] A. K. Moorthy and A.C. Bovik. LIVE wireless video quality assessment database. *http://live.ece.utexas.edu/research/quality/live_wireless_video.html*.

[181] A. K. Moorthy, W. S. Geisler, and A. C. Bovik. Evaluating the task dependence on eye movements for compressed videos. *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, January 2010.

[182] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik. Wireless video quality assessment: A study of subjective scores and objective algorithms'. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(4):513–516, April 2010.

[183] M. C. Morrone, M. Di Stefano, and D. C. Burr. Spatial and temporal properties of neurons of the lateral suprasylvian cortex of the cat. *Journal of neurophysiology*, 56(4):969–986, 1986.

[184] R. Muijs and I. Kirenko. A no-reference blocking artifact measure for adaptive video processing. *Proceedings of the 13th European Signal Processing Conference (EUSIPCO05)*, 2005.

[185] D. Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 66(3):241–251, 1992.

[186] M. Naccari, M. Tagliasacchi, F. Pereira, and S. Tubaro. No-reference modeling of the channel induced distortion at the decoder for H. 264/AVC video coding. In *Proceedings of the International Conference on Image Processing, San Diego, CA, USA*, 2008.

[187] M.J. Nadenau, S. Winkler, D. Alleysson, and M. Kunt. Human Vision Models for Perceptually Optimized Image Processing–A Review. *Proceedings of the IEEE*, 2000, 2000.

[188] N.D. Narvekar and L.J. Karam. A No-reference Perceptual Image Sharpness Metric based on a Cumulative Probability of Blur Detection. *1st International Workshop on Quality of Multimedia Experience (QoMEX)*, 2009.

[189] A. V. Nasonov and A. S. Krylov. Adaptive Image Deringing. In *Proceeding of Graphicon*, 2009.

[190] A. V. Nasonov and A. S. Krylov. Scale-space Method of Image Ringing Estimation. In *IEEE International Conference on Image Processing*, pages 2793–2796, 2009.

[191] Y. Nie and K.K. Ma. Adaptive rood pattern search for fast block-matching motion estimation. *IEEE Transactions on Image Processing*, 11(12):1442–1449, 2002.

[192] Y. Nie and K.K. Ma. Adaptive irregular pattern search with matching prejudgment for fast block-matching motion estimation. *IEEE Trans-*

actions on Circuits and Systems for Video Technology, 15(6):789–794, 2005.

[193] A. Nikitin, V. Solovyev, V. Khryashchev, and A. Priorov. Adaptive bilateral filter for JPEG2000 deringing. In *IEEE International Conference on Image Processing, Computer Vision and Pattern Recognition*, 2011.

[194] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. On the performance of human visual system based image quality assessment metric using wavelet domain. *Proc. SPIE Human Vision and Electronic Imaging XIII*, 6806, 2008.

[195] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Jnl. Spl. Top. Sign. Proc.*, 3(2):253–265, April 2009.

[196] A. Nosratinia. Enhancement of JPEG-compressed images by re-application of JPEG. *The Journal of VLSI Signal Processing*, 27(1):69–79, 2001.

[197] A. Nosratinia. Postprocessing of JPEG-2000 Images to Remove Compression Artifacts. *IEEE Signal Processing Letters*, 10(10):296–299, 2003.

[198] Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, International Telecommunications Union. *Std. ITU-T Rec. J. 144*, 2004.

[199] T. Oelbaum and K. Diepold. Building a reduced reference video quality metric with very low overhead using multivariate data analysis. *International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA 2007)*, July 2007.

[200] T. Oelbaum and K. Diepold. A reduced reference video quality metric for avc/h.264. *Proc. European Signal Processing Conference*, pages 1265–1269, September 2007.

[201] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int'l. J. Comp. Vis.*, 42(3):145–175, 2001.

[202] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems, 7: 333–339*, 1996.

[203] B. A. Olshausen and D. J. Field. How close are we to understanding V1? *Neural Computation*, 17(8):1665–1699, 2005.

[204] E. P. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, and L. Jiang. No-reference JPEG-2000 image quality metric. In *International Conference on Multimedia and Expo*, volume 1, pages 6–9, 2003.

[205] E.P. Ong, S. Wu, M.H. Loke, S. Rahardja, J. Tay, C.K. Tan, and L. Huang. Video quality monitoring of streamed videos. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1153–1156, 2009.

[206] A.V. Oppenheim and R.W. Schafer. *Discrete-time signal processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1989.

[207] J. Orchard, M. Ebrahimi, and A. Wong. Efficient nonlocal-means denoising using the svd. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1732–1735. IEEE.

[208] S.E. Palmer. *Vision science: Photons to phenomenology*. MIT press Cambridge, MA., 1999.

[209] H. Pan, X. F. Feng, and S. Daly. Lcd motion blur modeling and analysis. *IEEE International Conference on Image Processing*, pages II– 21–4, 2005.

[210] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik. Spatio-temporal quality pooling accounting for transient severe impairments and egomotion. *IEEE International Conference on Image Processing (ICIP)*, 2010.

[211] R.R. Pastrana-Vidal and J.C. Gicquel. Automatic quality assessment of video fluidity impairments using a no-reference metric. *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2006.

[212] R.R. Pastrana-Vidal and J.C. Gicquel. A no-reference video quality metric based on a human assessment model. In *Third International*

Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM, volume 7, pages 25–26, 2007.

[213] R.R. Pastrana-Vidal, J.C. Gicquel, C. Colomes, and H. Cherifi. Frame dropping effects on user quality perception. *5th International Workshop on Image Analysis for Multimedia Interactive Services*, 2004.

[214] R.R. Pastrana-Vidal, J.C. Gicquel, C. Colomes, and H. Cherifi. Sporadic frame dropping impact on quality perception. *Proceedings of SPIE*, 5292:182, 2004.

[215] PCWorld. FCC Warns of Impending Wireless Spectrum Shortage. *http://www.pcworld.com/article/186434/fcc_warns_of_impending_wireless_spectrum_shortage* 2010.

[216] H.A. Peterson, A.J. Ahumada Jr, and A.B. Watson. An improved detection model for DCT coefficient quantization. *Human Vision, Visual Processing, and Digital Display IV*, pages 191–201.

[217] M. H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. *Visual Communications and Image Processing, SPIE*, 5150:573582, 2003.

[218] M. H. Pinson and S. Wolf. The impact of monitor resolution and type on subjective video quality testing. *Technical Report, NTIA*, 2004.

[219] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, (3):312–313, September 2004.

[220] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti. Tampere image database. *http://www.ponomarenko.info/tid2008.htm*, 2008.

[221] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - A Database for Evaluation of Full Reference Visual Quality Assessment Metrics. *Advances of Modern Radioelectronics*, 10:30–45, 2009.

[222] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image Denoising using Scale Mixtures of Gaussians in the Wavelet Domain. *IEEE Transactions on Image Processing*, 12(11):1338–1351, 2003.

[223] H. Rabbani, M. Vafadoost, I. Selesnick, and S. Gazor. Image denoising based on a mixture of bivariate gaussian models in complex wavelet domain. In *IEEE International Summer School on Medical Devices and Biosensors*, pages 149–153, 2006.

[224] H. Rabbani, M. Vafadust, I. Selesnick, and S. Gazor. Image denoising employing a mixture of circular symmetric Laplacian models with local parameters in complex wavelet domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 805–808, 2007.

[225] R.P.N. Rao and D.H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *nature neuroscience*, 2:79–87, 1999.

[226] I.E.G. Richardson. H. 264 and MPEG-4 video compression. 2003.

[227] M. Ries, O. Nemethova, and M. Rupp. Performance evaluation of mobile video quality estimators. In *Proceedings of the European Signal Processing Conference,(Poznan, Poland*. Citeseer, 2007.

[228] K. Rijkse. H. 263: Video coding for low-bit-rate communication. *IEEE Communications Magazine*, 34(12):42–45, 1996.

[229] G. Roth, R. Sjoberg, G. Liebl, T. Stockhammer, V. Varsa, and M. Karczewicz. Common test conditions for rtp/ip over 3gpp/3gpp2. *ITU-T SG16 Doc. VCEG-M77*, 2001.

[230] D.M. Rouse and S.S. Hemami. Understanding and simplifying the structural similarity metric. *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, pages 1188–1191, 2008.

[231] D. L. Ruderman. The Statistics of Natural Images. *Network computation in neural systems*, 5(4):517–548, 1994.

[232] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[233] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How mt cells analyze the motion of visual patterns. *Nature Neuroscience*, 9(11):1421–1431, November 2006.

[234] M. A. Saad and A. C. Bovik. Natural motion statistics for no-reference video quality assessment. *International Workshop on Quality of Multimedia Experience*, pages 163–167, 2009.

[235] M. A. Saad, A. C. Bovik, and C. Charrier. A perceptual DCT Statistics based Blind Image Quality Metric. *IEEE Signal Processing Letters*, 17(6):583–586, 2010.

[236] N.G. Sadaka, L.J. Karam, R. Ferzli, and G.P. Abousleman. A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling. In *15th IEEE International Conference on Image Processing.*, pages 369–372, 2008.

[237] A. Said and W. A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circuits Syst. Video Technol.*, 6:243250, 1996.

[238] J. Salmon. On two parameters for denoising with non-local means. *IEEE Signal Processing Letters*, 17(3):269–272, 2010.

[239] J. Salmon and Y. Strozecki. From patches to pixels in non-local methods: Weighted-average reprojection. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1929–1932. IEEE, 2010.

[240] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Processing*, 18(11):2385–2401, October 2009.

[241] Sandvine. Global Internet Phenomena Spotlight. *http://www.sandvine.com/downloads/documents/05-17-2011_phenomena/Sandvine%20Global%20Internet%20Phenomena%20Spotlight%20-%20Netflix%20Rising.pdf*, 2011.

[242] Z. M. P. Sazzad, Y. Kawayoke, and Y. Horita. No reference image quality assessment for JPEG2000 based on spatial features. *Signal Processing: Image Communication*, 23(4):257–268, 2008.

[243] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

[244] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the h. 264/avc standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(9):1103–1120, 2007.

[245] R. Sekuler and R. Blake. *Perception*. Random House USA Inc, 1988.

[246] R. Sekuler and R. Blake. *Perception*. McGraw Hill, 2002.

[247] K. Seshadrinathan. *Video quality assessment based on motion models*. PhD thesis, The University of Texas at Austin, 2008.

[248] K. Seshadrinathan. MOVIE Software Release. *http://live.ece.utexas .edu/research/Quality/movie.html*, 2010.

[249] K. Seshadrinathan and A. C. Bovik. A structural similarity metric for video based on motion models. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, pages 869–872, April 2007.

[250] K. Seshadrinathan and A. C. Bovik. Unifying analysis of full reference image quality assessment. *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, pages 1200–1203, 2008.

[251] K. Seshadrinathan and A. C. Bovik. *The Essential Guide to Video Processing*, chapter Video Quality Assessment. Academic Press, 2009.

[252] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on*, 19(2):335–350, 2010.

[253] K. Seshadrinathan and A.C. Bovik. Temporal hysteresis model of time varying subjective video quality. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 1153–1156. IEEE, 2011.

[254] K. Seshadrinathan, R. J. Safranek, J. Chen, T. N. Pappas, H. R. Sheikh, E. P. Simoncelli, Z. Wang, and A. C. Bovik. *Image quality assessment in*

The Essential Guide to Image Processing, chapter 20. Academic Press, 2009.

[255] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(2):1427–1441, 2010.

[256] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L. K. Cormack. LIVE video quality assessment database. *http://live.ece.utexas.edu/research/quality/live_video.html*.

[257] J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Process*, 41(12):34453462, 1993.

[258] K. Sharifi and A. Leon-Garcia. Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. *IEEE Tran. Circ. Syst. for Video Tech.*, 5(1):52–56, 1995.

[259] G Sharma. Lcds versus crts: Color-calibration and gamut considerations. *Proceedings of the IEEE*, 90(4):605–622, April 2002.

[260] H. R. Sheikh and A. C. Bovik. A visual information fidelity approach to video quality assessment. *First International Workshop on Video Processing and Quality Metrics for Conusmer Electronics*, January 2005.

[261] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

[262] H. R. Sheikh, A. C. Bovik, and L. K. Cormack. No-reference quality assessment using natural scene statistics: JPEG 2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005.

[263] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005.

[264] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, November 2006.

[265] H.R. Sheikh, Z. Wang, L. Cormack, and A.C. Bovik. LIVE image quality assessment database. *2007-01-20]. http://live. ece. utexas, edu/researeh/quality.*

[266] D. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. CRC Press, 2004.

[267] A. Shnayderman, A. Gusev, and A.M. Eskicioglu. A multidimensional image quality measure using singular value decomposition. 5294:82–92, 2003.

[268] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992.

[269] E. P. Simoncelli and B. A. Olshausen. Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.

[270] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S. C. Zhu. On advances in statistical modeling of natural images. *J. Math. Imaging Vis.*, 18(1):17–33, 2003.

[271] J. L. Starck, E. J. Candes, and D. L. Donoho. The curvelet transform for image denoising. *IEEE Transactions on Image Processing*, 11(6):670–684, 2002.

[272] A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4):578–585, 2006.

[273] T. Stockhammer, M.M. Hannuksela, and T. Wiegand. H.264/avc in wireless environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):657–673, 2003.

[274] A. Stuart and J. K. Ord. The advanced theory of statistics. 1977.

[275] O. Sugimoto, R. Kawada, M. Wada, and S. Matsumoto. Objective measurement scheme for perceived picture quality degradation caused by MPEG encoding without any reference pictures. *Proceedings of SPIE*, 4310:932, 2000.

[276] D. Sun and W. K. Cham. An Effective Postprocessing Method for Low Bit Rate Block DCT Coded Images. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 705–708, 2007.

[277] D. Sun and W. K. Cham. Postprocessing of Low Bit-Rate Block DCT Coded Images based on a Fields of Experts Prior. *IEEE Transactions on Image Processing*, 16(11):2743–2751, 2007.

[278] S. Suthaharan. Perceptual quality metric for digital video coding. *Electronics Letters*, 39(5):431–433, 2003.

[279] S. Suthaharan. No-reference visually significant blocking artifact metric for natural scene images. *Signal Processing*, 89(8):1647–1652, 2009.

[280] H. Takeda, S. Farsiu, and P. Milanfar. Kernel regression for image processing and reconstruction. *Transactions on Image Processing*, 16:349–366, 2007.

[281] K. T. Tan and M. Ghanbari. Blockiness detection for MPEG2-coded video. *IEEE Signal Processing Letters*, 7(8):213–215, 2000.

[282] D. S. Taubman and M. W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.

[283] P.C. Teo and D.J. Heeger. Perceptual image distortion. *SID International Symposium Digest of Technical Papers*, 25:209–209, 1994.

[284] D.J. Tolhurst and J.A. Movshon. Spatial and temporal contrast sensitivity of striate cortical neurones. *Nature*, 257:674–675, 1975.

[285] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *International Conference on Computer Vision*, pages 839–846. IEEE, 1998.

[286] H. Tong, M. Li, H. J. Zhang, and C. Zhang. No-reference quality assessment for JPEG2000 compressed images. In *IEEE Intl. Conf. Img. Proc.*, pages 24–27. Citeseer, 2004.

[287] S. Tourancheau, P. Le Callet, and D. Barba. Impact of the resolution on the difference of perceptual video quality between crt and lcd. *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 3:III–441–III–444, 2007.

[288] International Telecommuncation Union. Bt-500-11: Methodology for the subjective assessment of the quality of television pictures. international telecommuncation union. *Recommendation*.

[289] G. Valenzise, M. Naccari, M. Tagliasacchi, and S. Tubaro. Reduced-reference estimation of channel-induced video distortion using distributed source coding. *Proceeding of the 16th ACM international conference on Multimedia*, 2008.

[290] D. Van De Ville and M. Kocher. Sure-based non-local means. *IEEE Signal Processing Letters*, 16(11):973–976, 2009.

[291] D. Van De Ville and M. Kocher. Non-local means with dimensionality reduction and sure-based parameter selection. *Transactions on Image Processing*, 20:2683 – 2690, 2011.

[292] C.J. Van den Branden Lambrecht and O. Verscheure. Perceptual quality measure using a spatiotemporal model of the human visual system. *Proceedings of SPIE.*, pages 450–461, 1996.

[293] C.J. Van den Branden Lambrecht and O. Verscheure. Perceptual quality measure using a spatiotemporal model of the human visual system. *Proc. SPIE*, pages 450–461, 1996.

[294] A. M. van Dijk, J. B. Martens, and A. B. Watson. Quality asessment of coded images using numerical category scaling. *SPIE Advanced Image and Video Communications and Storage Technologies*, 2451:90–101, 1995.

[295] V.N. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, 2000.

[296] S. Varadarajan and L.J. Karam. An improved perception-based no-reference objective image sharpness metric using iterative edge refinement. In *Proceedings of the 15th IEEE International Conference on Image Processing*, pages 401–404, 2008.

[297] Video Quality Experts Group (VQEG). Final report from the video quality experts group on the validation of objective quality metrics for

video quality assessment phase I. *http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI*, 2000.

[298] Video Quality Experts Group (VQEG). Final report of video quality experts group multimedia phase I validation test, TD 923, ITU Study Group 9. 2008.

[299] T. Vlachos. Detection of blocking artifacts in compressed video. *Electronics Letters*, 36(13):1106–1108, 2000.

[300] Video Quality Experts Group (VQEG). Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment phase II. *http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII*, 2003.

[301] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli. Natural image statistics and divisive normalization: modeling nonlinearities and adaptation in cortical neurons. *Statistical theories of the brain*, pages 203–222, 2002.

[302] M.J. Wainwright and E.P. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. *Advances in neural information processing systems*, 12(1):855–861, 2000.

[303] B.A. Wandell. *Foundations of vision*. Sinauer Associates, 1995.

[304] G. Wang, T. T. Wong, and P. A. Heng. Deringing cartoons by image analogies. *ACM Transactions on Graphics*, 25(4):1360–1379, 2006.

[305] J. Wang, Y. Guo, Y. Ying, Y. Liu, and Q. Peng. Fast non-local algorithm for image denoising. In *Image Processing, 2006 IEEE International Conference on*, pages 1429–1432. IEEE, 2006.

[306] T. Wang and G. Zhai. JPEG2000 image postprocessing with novel trilateral deringing filter. *Optical Engineering*, 47:027005, 2008.

[307] Z. Wang. The structural similarity index. *http://live.ece.utexas.edu/research/Quality/index.htm*.

[308] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002.

[309] Z. Wang and A. C. Bovik. *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006.

[310] Z. Wang and A. C. Bovik. Mean squared error: love it or leave it? - a new look at signal fidelity measures. *IEEE Signal Proceessing Magazine*, 26(1):98–117, 2009.

[311] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error measurement to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[312] Z. Wang, A.C. Bovik, and B.L. Evans. Blind measurement of blocking artifacts in images. In *Proc. IEEE Int. Conf. Image Proc*, volume 3, pages 981–984. Citeseer, 2000.

[313] Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *Jnl. Opt. Soc. Am.*, 24(12):B61–B69, December 2007.

[314] Z. Wang, L. Lu, and A. C. Bovik. Video quality assesssment based on structural distortion measurement. *Signal Processing: Image communication*, (2):121–132, February 2004.

[315] Z. Wang and X. Shang. Spatial pooling strategies for perceptual image quality assessment. *IEEE international conference on Image Processing*, September 2006.

[316] Z. Wang, H. R. Sheikh, and A. C. Bovik. No-reference perceptual quality assessment of jpeg compressed images. *Proc. of IEEE Int. Conf. on Image Processing*, 1:477–480, 2002.

[317] Z. Wang, H.R. Sheikh, and A.C. Bovik. Objective video quality assessment. *The Handbook of Video Databases: Design and Applications*, pages 1041–1078, 2003.

[318] Z. Wang and E. P. Simoncelli. Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):1–13, September 2008.

[319] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for image quality assessment. *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar)*, November 2003.

[320] Z. Wang and EP Simoncelli. Translation insensitive image similarity in complex wavelet domain. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, 2, 2005.

[321] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.H. Yang, and A. C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15(6):1680–1689, 2006.

[322] A. B. Watson. The cortex transform- Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing*, 39(3):311–327, 1987.

[323] A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of the Optical Society of America A*, 2(2):322–342, 1985.

[324] A.B. Watson, J. Hu, and J.F. McGowan III. Digital video quality metric based on human vision. *Jnl. Ele. Imag.*, 10:20, 2001.

[325] T. Wiegand, GJ Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003.

[326] S. Winkler. A perceptual distortion metric for digital color video. *Proc. SPIE*, 3644(1):175–184, 1999.

[327] S. Winkler and R. Campos. Video quality evaluation for Internet streaming applications. *Proceedings of SPIE Human Vision and Electronic Imaging*, 5007:21–24, 2003.

[328] S. Winkler and F. Dufaux. Video quality evaluation for mobile applications. In *Proc. of SPIE Conference on Visual Communications and Image Processing, Lugano, Switzerland*, volume 5150, pages 593–603. Citeseer, 2003.

[329] S. Winkler, A. Sharma, and D. McNally. Perceptual video quality and blockiness metrics for multimedia streaming applications. In *Proceedings of the International Symposium on Wireless Personal Multimedia Communications*, pages 547–552, 2001.

[330] S. Wolf and M. Pinson. Video quality measurement techniques. *National Telecommunications and Information Administration (NTIA) Report 02-392*, 2002.

[331] S. Wolf and M.H. Pinson. Low bandwidth reduced reference video quality monitoring system. *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*, 2005.

[332] T. Yamada, Y. Miyamoto, and M. Serizawa. No-reference video quality estimation based on error-concealment effectiveness. *Packet Video 2007*, pages 288–293, 2007.

[333] T. Yamada, Y. Miyamoto, M. Serizawa, and H. Harasaki. Reduced-reference based video quality-metrics using representative-luminance values. *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, Scottsdale, AZ, USA*, 2007.

[334] F. Yang, S. Wan, Y. Chang, and H.R. Wu. A novel objective no-reference metric for digital video quality assessment. *IEEE Signal processing letters*, 12(10):685–688, 2005.

[335] K.C. Yang, C. C. Guest, K. El-Maleh, and P. K. Das. Perceptual temporal quality metric for compressed video. *IEEE Transactions on Multimedia*, 9(7):1528–1535, 2007.

[336] S. Yao, W. Lin, Z. Lu, E. Ong, M. Locke, and S. Wu. Image quality measure using curvature similarity. *IEEE International Conference on Image Processing*, pages 437–440, 2007.

[337] C. Yim and A. C. Bovik. Quality assessment of deblocked images. *IEEE Transactions on Image Processing*, 20(1):88–98, 2011.

[338] M. Yuen and HR Wu. A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing*, 70(3):247–278, 1998.

[339] W.L. Zeng and X.B. Lu. Region-based non-local means algorithm for noise removal. *Electronics Letters*, 47(20):1125–1127, 2011.

[340] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang. Image deringing using quadtree based block-shift filtering. In *International Symposium on Circuits and Systems*, pages 708–711. IEEE, 2008.

[341] G. Zhai, W. Zhang, X. Yang, W. Lin, and Y. Xu. Efficient image deblocking based on postfiltering in shifted windows. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(1):122–126, 2008.

[342] R. Zhang, Y. L. Fong, and W. K. Cham. Image deblocking by the dual adaptive FIR wiener filter and overcomplete representation. In *IEEE International Conference on Information, Communications and Signal Processing*, pages 1–4, 2009.

[343] S. Zhu and K.K. Ma. A new diamond search algorithm for fast block-matching motionestimation. *IEEE Transactions on Image Processing*, 9(2):287–290, 2000.

[344] X. Zhu and P. Milanfar. A no-reference sharpness metric sensitive to blur and noise. In *1st International Workshop on Quality of Multimedia Experience (QoMEX)*, 2009.

[345] X. Zhu and P. Milanfar. Automatic parameter selection for denoising algorithms using a no-reference measure of image content. *IEEE Transactions on Image Processing*, 19(12):3116–3132, 2010.